

AI in control: Rethinking cybersecurity compliance and auditing[☆]

Fatma Yasmine Loumachi ^a, Marcio J. Lacerda ^a,^{*}, Karim Ouazzane ^a, Asma Adnane ^b,
Oksana Adamyk ^b

^a School of Computing and Digital Media, London Metropolitan University, London, UK

^b Loughborough University, Loughborough, UK

ARTICLE INFO

Keywords:

Cybersecurity compliance
Audit assurance
Artificial intelligence
Epistemic reasoning
Meta-compliance
Symbolic governance

ABSTRACT

Context: Placing Artificial Intelligence (AI) in control of cybersecurity compliance and auditing shifts its role from decision-support to direct execution of regulatory operational processes, where AI outputs may constitute compliance artefacts and audit evidence. This raises the problem of *Meta-Compliance*, in which not only the organisation but also the AI system must satisfy enforceable requirements. Yet existing frameworks provide no operational criteria for recognising AI as authoritative in such roles. Trustworthy AI principles define high-level Second-Layer requirements but remain non-binding, whereas First-Layer organisational requirements impose explicit justificatory and evidentiary duties.

Objectives: This study investigates the minimal normative conditions under which AI systems can be recognised as authoritative in compliance and auditing, capable of producing evidence valid for assurance.

Methods: Doctrinal analysis is conducted on binding “shall/must” provisions across PCI DSS, DORA, UK GDPR, NIS2, ISO/IEC 27001, and NIST SP 800-53. Provisions are normalised through the compliance–audit chain (*requirement* → *control* → *rule* → *evidence*) and mapped against Second-Layer AI governance requirements. The result is the *Compliance–Audit Authority Benchmark (CAAB)*, comprising six criteria: Traceability, Explainability, Evidence Integrity, Adaptability, Action Governance, and Reasoning.

Results: Applying CAAB across AI model families and architectures shows that symbolic and knowledge-representation methods satisfy most criteria intrinsically, whilst neural, deep, and generative models do not unless supported by external governance mechanisms. This exposes a structural gap between First-Layer organisational requirements and Second-Layer AI requirements, clarifying that authority rests on evidentiary guarantees rather than statistical accuracy.

Conclusion: The study formalises *Meta-Compliance* as the recursive structure in which both organisations and AI systems become subjects of assurance. CAAB defines the minimum conditions for recognising AI as authoritative, whilst the proposed *Verifiable Reasoning Architecture (VRA)* may offer a pathway towards AI systems anchored in secured evidence, reproducible inference, and symbolic governance, establishing audit-ready authority in high-risk contexts.

Contents

1.	Introduction and background	2
1.1.	Compliance and audit practices	2
1.2.	Regulations and standards	2
1.3.	AI in compliance and auditing	3
1.4.	Control and authority	3
1.5.	A two-layered compliance structure	3
1.6.	Gap and research question	3
1.7.	Analytical approach	3
1.8.	Contributions	3

[☆] This article is part of a Special issue entitled: ‘RegCompliance in SE’ published in Information and Software Technology.

^{*} Corresponding author.

E-mail addresses: fal0167@my.londonmet.ac.uk (F.Y. Loumachi), m.lacerda@londonmet.ac.uk (M.J. Lacerda), k.ouazzane@londonmet.ac.uk (K. Ouazzane), a.adnane@lboro.ac.uk (A. Adnane), o.o.adamyk@lboro.ac.uk (O. Adamyk).

<https://doi.org/10.1016/j.infsof.2026.108132>

Received 6 October 2025; Received in revised form 11 February 2026; Accepted 18 March 2026

Available online 28 March 2026

0950-5849/© 2026 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).

2.	First-layer compliance	4
2.1.	Reference model.....	4
3.	Second-layer compliance	4
3.1.	Candidate AI requirements	5
3.2.	AI for cybersecurity	5
4.	Methodology	6
4.1.	Corpus delimitation	6
4.2.	Coding and interpretation	6
4.3.	Derivation of criteria	6
4.4.	Demonstrator	9
4.4.1.	CAAB × AI taxonomy (families).....	9
4.4.2.	CAAB × reference model.....	9
5.	Meta-compliance.....	9
5.1.	Compliance-Audit Authority Benchmark (CAAB)	9
5.2.	CAAB × AI taxonomy (families).....	9
5.3.	CAAB × reference model.....	10
6.	Evaluation of findings	11
6.1.	Evaluation of CAAB × reference model	11
6.2.	Evaluation of CAAB × AI taxonomy	11
6.3.	Datasets and evaluation methods.....	11
6.4.	Discussion.....	15
6.4.1.	Implications for system design.....	15
6.4.2.	Implications for governance	16
7.	Systemic challenges, regulatory constraints, and research gaps	16
7.1.	Challenges and constraints	16
7.1.1.	Opaque systems in a rule-based domain.....	16
7.1.2.	Control inference without control enforcement	16
7.1.3.	From risk reduction to risk obfuscation.....	16
7.1.4.	Operational security by cosmetic compliance.....	16
7.1.5.	Audit evidence as a trust proxy	16
7.2.	Identified gaps in current practices	16
7.2.1.	Fragility of single-model AI dependence.....	16
7.2.2.	Lack of provenance	16
7.2.3.	Epistemic misalignment with regulatory logic.....	16
7.2.4.	Undefined accountability in audit chains.....	16
7.3.	Answer to the research question	17
8.	Conceptual solution, future direction and limitations	17
8.1.	Formal specification (minimal)	17
8.2.	Axioms.....	17
8.3.	Limitations	17
9.	Conclusion	18
	CRediT authorship contribution statement	18
	Declaration of competing interest.....	18
	Data availability	18
	References.....	18

1. Introduction and background

Cybersecurity compliance and auditing are core mechanisms of governance and accountability across organisational and transnational systems [1]. Compliance denotes adherence to applicable regulations and contractual standards through the implementation of mandated controls and the retention of evidentiary artefacts that attest to their execution, including logs, configurations, scan outputs, and change records [1–5].

Auditing, whilst distinct, remains tightly coupled with compliance as an evidence-based process for validating conformance. It assesses whether controls and associated processes meet stated requirements and whether accumulated evidence is sufficient and appropriate for assurance [3,6,7]. In organisational practice, compliance activities and evidence production occur continuously, whilst audits take place at discrete intervals, either internally or externally [4,6,7].

Between abstract requirements and concrete evidence sit operationalised assessment rules. These rules encode interpretations of control intent and define enforceable evaluation criteria. Requirements, controls, rules, and evidence form an operational chain through which cybersecurity requirements are realised in practice [4,5].

1.1. Compliance and audit practices

The epistemology of cybersecurity compliance and auditing has historically been dominated by reactive, checklist-based validation. Organisational compliance efforts emphasise formal demonstrability, focusing on documented requirements mapped to nominal controls, instead of continuous assessment of control effectiveness under evolving real-world threat conditions [8].

Auditing practices reinforce this epistemic orientation. Audits commonly rely on ex post, sample-based inspection and human interpretation of accumulated artefacts [9]. Assurance conclusions are frequently indirect and contingent on abstraction and professional judgement. Regulatory assurance exhibits weak alignment with continuous control effectiveness, runtime system behaviour, and adversarial resilience [3, 6].

1.2. Regulations and standards

Within this context, cybersecurity regulations and standards define compliance requirements with differing scopes, legal force, and enforcement mechanisms. The Payment Card Industry Data Security Standard (PCI DSS) [3], for example, constitutes an industry-defined

standard governing the protection of cardholder data. By contrast, the General Data Protection Regulation (GDPR) [10] is a statutory regulation that imposes legally binding requirements on the processing and protection of personal data.

Regulations are enacted by public authorities and impose enforceable requirements on organisations within scope. Standards, whilst not laws in themselves, typically originate from industry consortia or technical bodies and acquire binding force through contractual commitments, certification schemes, or explicit regulatory incorporation [3]. Despite these structural differences, regulatory and standards-based requirements converge at the operational level, particularly within processes such as risk assessment (RA) and incident response (IR). This convergence embeds security requirements into repeatable organisational practices [11].

1.3. AI in compliance and auditing

Artificial intelligence (AI) now increasingly intersects with the compliance–audit chain. Manual certification processes and checklist-based inspections are supplemented, and in some cases replaced, by real-time monitoring and algorithmic evaluation [12]. Penetration testing (PT) is extended through automated red-teaming approaches [13], whilst anomaly detection increasingly relies on machine learning (ML) techniques [14].

These developments place AI in substantive compliance and auditing roles beyond decision support.

1.4. Control and authority

When AI systems operate directly within compliance and auditing processes, assurance extends beyond organisational conformance with cybersecurity requirements to include the systems performing compliance checks. We formalise this extension as *Meta-Compliance*, a second-order assurance problem. This structure brings forward the twin notions of control and authority. Placing AI in control involves delegating to it the ability to execute, select, or veto compliance and audit processes without immediate human validation [15]. Authority is distinct. In organisational theory, it denotes the recognised right to issue determinations that bind outcomes and are accepted as valid by oversight actors [16]. Applied to AI, authority implies that system outputs may constitute binding compliance artefacts or audit evidence [17, 18]. In high-risk cybersecurity contexts, this raises operational and legal challenges [15].

1.5. A two-layered compliance structure

When AI systems participate directly in compliance and auditing operations, compliance requirements extend beyond organisational activities to the mechanisms performing compliance determinations. This gives rise to two distinct loci of compliance. One concerns the organisation and its regulated operations. The other concerns the AI systems entrusted with executing or supporting compliance and audit functions.

We refer to the former as First-Layer compliance, which pertains to organisational conformance with cybersecurity regulations and standards. The latter is termed Second-Layer compliance and pertains to requirements applicable to AI systems performing compliance and audit functions, aligned with principles of Trustworthy AI [19].

Trustworthy AI provides a basis for assessing whether AI systems operating in high-risk contexts can be relied upon to act in lawful, ethical, and technically robust ways. It also serves as a reference frame for aligning requirements across AI governance frameworks and international standards [20].

1.6. Gap and research question

However, existing requirements do not resolve the practical challenges associated with AI deployment in cybersecurity compliance and auditing. Organisations must demonstrate not only that monitoring and detection functions occur, but also that the AI systems performing these functions operate in a compliant, reliable, and auditable manner, with outputs recognised as valid audit evidence [21].

Viewed through a Meta-Compliance lens, a structural gap becomes apparent. First-Layer compliance requirements provide limited guidance on the permissibility and conditions of AI-performed compliance and audit functions. Second-Layer compliance requirements articulate high-level principles but lack operational criteria for implementation within AI-based compliance systems and tooling.

From this gap arises the central research question of this study: if AI systems are entrusted with cybersecurity compliance and auditing, under what criteria can they be recognised as authoritative, such that their determinations are accepted as valid compliance artefacts and audit evidence? Despite increasing reliance on AI within compliance systems and tooling, no binding regulatory framework or standard specifies the evidentiary and reasoning conditions required for such recognition.

1.7. Analytical approach

Addressing this gap necessitates doctrinal analysis of binding regulatory provisions governing compliance requirements, controls, and evidence. Cybersecurity compliance and auditing are grounded in normative frameworks that define valid conformity, evidentiary sufficiency, and acceptable assurance, and these determinations are assessed through established regulatory and audit practices.

When AI systems assume control or decision roles within compliance and auditing processes, the evaluation of their outputs follows the same normative criteria applied to organisational compliance. This reflects how regulatory and audit judgement establishes validity through conformity with prescribed requirements, control expectations, and evidentiary standards.

Doctrinal analysis supports reconstruction of the *requirement* → *control* → *rule* → *evidence* chain of First-Layer compliance and its mapping to corresponding Second-Layer compliance requirements applicable to AI systems. From this mapping, criteria for authority recognition are derived and applied to AI architectures and systems in a non-empirical, evaluative manner.

1.8. Contributions

Accordingly, this paper makes the following contributions:

- It conducts a doctrinal analysis of binding cybersecurity provisions and maps these against AI governance requirements to derive and normalise the Compliance–Audit Authority Benchmark (CAAB) as a set of prerequisite criteria.
- It applies CAAB to selected AI model families and systems to identify systematic gaps in existing approaches.
- It proposes a high-level Verifiable Reasoning Architecture (VRA) as a design pathway aligned with the conditions identified by CAAB.

Scope note

The criteria derived in this study are not presented as determinative or sufficient for authority in any specific institutional setting. Rather, they specify minimum conditions that must be satisfied before AI-generated outputs can, in principle, be considered eligible for recognition as compliance artefacts or audit evidence under existing

regulatory and audit practices. Application of these criteria to AI architectures does not constitute enforcement, certification, or system validation. Forensic and jurisdiction-specific sufficiency, audit determinations, liability, and legal admissibility fall outside the scope of this study.

2. First-layer compliance

As previously defined, First-Layer compliance refers to organisational adherence to cybersecurity regulations and standards. This section establishes the analytical basis of the study by examining how compliance requirements are operationalised across multiple regulations and standards and how these operational patterns recur in practice. The resulting analysis identifies a set of operational processes that provide the reference frame for mapping Second-Layer requirements and assessing AI systems.

2.1. Reference model

The reference model is constructed in this study through clause-level analysis of binding requirements, abstracting recurring operational actions into a minimal set of processes.

The model is grounded in a corpus of regulations and standards imposing mandatory cybersecurity requirements across organisational contexts in the UK, EU, and U.S. It covers DORA, GDPR/UK GDPR, and NIS2 as statutory cybersecurity and data protection regulations; PCI DSS v4.0.1 as a contractually mandatory industry standard in payment ecosystems; and ISO/IEC 27001:2022 and NIST SP 800-53 Rev. 5 as de facto organisational security standards referenced in regulatory guidance, audit practice, and assurance engagements.

Recurring requirement markers that signal operational activity (e.g., monitor, detect, log, review, retain, report, implement) were identified across the corpus and analysed within the *requirement* → *control* → *rule* → *evidence* chain. These markers were then consolidated into a minimal set of high-level operational processes that recur across the corpus and correspond to stages of the compliance and audit lifecycle where AI systems are practically deployed or proposed. The resulting processes were cross-mapped against a survey of major cybersecurity compliance frameworks by Wang et al. [22] to support coverage validation against commonly recognised operational activities. Table 1 presents the resulting First-Layer operational abstraction and the marker families and abstraction logic used to form each process. These processes define the operational points at which AI systems may participate in, or assume control over, compliance and audit activities in later sections.

Operational processes are instantiated through the chain *requirement* → *control* → *rule* → *evidence*. IR illustrates this recurrence across diverse regulations and standards. PCI DSS v4.0.1 requires the maintenance of an IR plan (e.g., Requirement 12.10), GDPR/UK GDPR mandates breach notification within seventy-two hours (Article 33 GDPR), NIS2 requires reporting of significant incidents (Article 23), and DORA specifies conditions for ICT incident management (Article 17). Despite differences in wording and scope, these requirements converge on the same operational process [23]. In practice, IR is realised through documented procedures and trained teams, articulated in rules such as escalation workflows and notification timeframes, and evidenced through artefacts including incident logs, post-incident reports, and regulator notifications [5].

The main operational processes of the reference model are presented below:

- **Vulnerability and Risk Assessments (VA-RA):** Systematic identification and evaluation of vulnerabilities and risks in in-scope systems and configurations. The process assesses severity and impact, contextualises findings against assets and threat conditions, and determines risk treatment priorities. Outputs support remediation decisions, risk acceptance, and prioritisation, and generate structured artefacts for compliance justification and audit review.

- **Requirement-to-Control Mapping and Gap Analysis:** Validates that compliance requirements are realised through defined technical and procedural controls and that gaps between requirements and implemented measures are identified and prioritised. The process establishes explicit traceability between requirements, controls, and supporting artefacts, producing auditable mappings that must be maintained as regulations, standards, and organisational practices evolve.
- **Control Implementation Validation:** Verifies that deployed security controls operate as intended and meet defined compliance and security objectives. Validation includes static methods such as configuration reviews and compliance scanning, and dynamic methods such as PT and red-team exercises. The process produces evidence-based results suitable for regulatory and audit review.
- **Evidence Collection and Management:** Governs how compliance and audit artefacts are identified, collected, retained, and assessed to meet requirements of sufficiency, appropriateness, and integrity. The process ensures that evidence supporting compliance determinations is verifiable, provenance-aware, and suitable for independent review.
- **Continuous Monitoring and Anomaly Detection:** Ongoing observation and analysis of system, network, and application activity to identify deviations from established normal behaviour [24, 25]. The process provides continuous coverage of in-scope assets, adapts to changing baselines and threat conditions, and generates artefacts supporting detection, investigation, and audit review.
- **Incident Response (IR) and Recovery:** Coordinates detection, analysis, containment, and remediation of security incidents whilst preserving system and evidentiary integrity. The process defines authorisation, execution, and documentation of response and recovery actions and produces verifiable records supporting post-incident analysis and control updates.
- **Audit and Reporting:** Assesses the sufficiency and effectiveness of implemented controls by reviewing evidence against defined objectives and issuing assurance reports for regulatory and independent review. Audit artefacts are retained through mechanisms that preserve integrity and chain of custody, ensuring conclusions are interpretable and defensible.

We identify seven operational processes in the reference model, defining the operational scope of the study and structuring the doctrinal mapping and evaluation. General organisational governance unrelated to cybersecurity control duties or audit-evidence conditions is excluded. Governance duties constraining authority, approvals, escalation, segregation, and accountability records are retained and mapped to the processes they condition. Audit is included, as evidentiary verification recurs across the analysed regulations and standards, despite differing conceptions; PCI DSS embeds audit as mandatory, whereas ISO/IEC 19011 [6] treats it as a distinct discipline.

With the First-Layer operational frame fixed, the analysis proceeds to Second-Layer compliance to examine AI-related requirements and situate AI models within the reference model.

3. Second-layer compliance

Second-Layer compliance refers to adherence to AI requirements grounded in the principles of Trustworthy AI [19]. In this section, these requirements are presented in their official form and aligned with First-Layer compliance to identify which can be treated as candidates for enforceable use. The analysis then turns to AI models in cybersecurity to determine which model families are historically recognised as operating within the operational processes represented in the First-Layer reference model and which can therefore be regarded as viable candidates in compliance and audit contexts.

Table 1
Abstraction of reference model operational processes from cybersecurity compliance regulations and standards.

Reference model operational process	Source keywords (PCI DSS, NIST, ISO, DORA, GDPR, NIS2)	Abstraction logic
Vulnerability and Risk Assessments (VA-RA)	Risk analysis, assessment, ranking, threat evaluation, justification, frequency	Systematic evaluation of risk/vulnerability
Requirement-to-control mapping & gap analysis	Policy, procedure, role, responsibility, requirement, governance, approval, exception (approved)	Obligations and governance structures
Control implementation validation	Implement, configure, deploy, enable, enforce, restrict, secure, encrypt, prevent, block, remove, isolate	Technical/procedural enforcement of requirements
Continuous monitoring and anomaly detection	Monitor, alert, detect, anomaly, behavioural, IDS, IPS, FIM, telemetry, tamper	Operational detection and monitoring processes
Evidence collection and management	Log, record, inventory, list, report, attestation, retention, documentation (artefacts), acknowledgement	Admissible artefacts required for proof of compliance
Incident Response (IR) and recovery	incident, exception (operational), anomaly addressed, containment, remediation, recovery, lessons learned	Response and restoration after abnormal or malicious events
Audit and reporting	review, verify, confirm, conform, test, inspect, audit, independent, oversight, sign-off	Periodic checks of control effectiveness

Indicative roles: *VA-RA*: security engineering, risk owners (management oversight); *Req-Control*: compliance functions, control owners; *Validation*: security engineering, platform/operations, independent testing; *Monitoring*: SOC, operations; *Evidence*: control owners, GRC/assurance; *IR*: CSIRT, operations (authorised); *Audit*: internal audit, external assessors, management sign-off.

Table 2
Screening of trustworthy AI requirements against binding First-Layer compliance duties.

Trustworthy AI requirement	Alignment with first-layer compliance (Operational analogue)	Decision
Traceability	Recurs as binding duties for documentation, scope definition, record retention, and demonstrable provenance across the compliance chain (<i>requirement</i> → <i>control</i> → <i>rule</i> → <i>evidence</i>).	Retain
Explainability	Recurs as binding expectations for reviewable determinations: documented criteria and scope, intelligible reporting, outcome justification, and disclosure of applied rules or thresholds for audit scrutiny.	Retain
Adaptability	Operationalised as binding change governance: periodic review, update, post-incident revision, and controlled modification preserving evidentiary continuity.	Retain
Auditability	Maps to binding evidence duties for preservation, integrity protection, tamper detection, authenticated logging, retention, and audit trails.	Retain
Fairness	No recurring binding operational analogue appears in the analysed corpus that bears directly on the admissibility of control evidence. The concept is central to human-impact decision regimes and would require separate doctrinal mapping (Deane J in Bhat [26]), outside the scope of this study.	Exclude (out of scope for this corpus)
Privacy	Binding where personal data is processed as a bright-line legality constraint; unlawful handling can invalidate evidence and preclude authoritative use. Treated as an eligibility condition, not a differentiating authority criterion.	Scope as gating constraint
Robustness	Expressed throughout the corpus as baseline security and resilience obligations. AI-specific failure modes are addressed through binding mechanisms for change control, evidentiary integrity, action governance, and reproducible inference, with no separate authority status in the corpus.	Scope as baseline; evaluated via CAAB
Societal well-being	Not expressed as enforceable operational duties in the analysed corpus and not linked to evidentiary admissibility in cybersecurity control assessment.	Exclude (no binding operational analogue)

3.1. Candidate AI requirements

Principles of Trustworthy AI are articulated in authoritative sources such as the Ethics Guidelines for Trustworthy AI and the NIST AI Risk Management Framework. In this study, these requirements are screened using the same doctrinal method applied to the First-Layer corpus: binding “shall/must” provisions are coded through the compliance chain *requirement* → *control* → *rule* → *evidence* and used to test whether each Trustworthy-AI requirement has a recurring operational analogue that functions as an evidence-condition for compliance determination.

A requirement is retained as a candidate authority criterion only where it recurs as binding duties that directly condition evidential admissibility (e.g., documentation and mapping, reviewable justification, integrity protection, change governance, retention, and audit trails). Requirements that operate as legality constraints in specific regimes

are treated as gating eligibility conditions, and requirements lacking a recurring binding operational analogue in the analysed cybersecurity compliance corpus are excluded at this stage. Table 2 reports the resulting candidate set used in the subsequent doctrinal mapping.

Having delimited enforceable requirements, the analysis proceeds to AI families relevant to compliance and audit contexts.

3.2. AI for cybersecurity

AI requirements gain practical relevance only when expressed in operational forms that systems can implement [34]. The scope is therefore limited to AI in cybersecurity, where AI systems already operate, or are proposed to operate, within the compliance and audit processes defined in the reference model.

Table 3
Selected surveys on AI in cybersecurity.

Study	Scope and description
Ferrag et al. [27]	Reviews LLM and generative AI uses in cybersecurity (2023–2024), including forensics, anomaly detection, and threat intelligence, with limits and mitigation strategies.
Kheddar [28]	Examines transformer/LLM models for IDS, assessing reliability, interpretability, scalability, and adaptability under evolving threats.
Kheddar et al. [29]	Surveys reinforcement approaches for IDS, especially DDoS; highlights adaptability and anomaly detection strengths versus training/data weaknesses.
Maniriho et al. [30]	Reviews classical ML, DL, sequential models, and DRL in malware detection across desktop and mobile.
Sikos [31]	Focuses on knowledge graphs, ontologies, and reasoning for traceability, explainability, and adaptive analysis; notes gaps in hybrid neuro-symbolic ML.
Rjoub et al. [32]	Defines explainable AI in security; maps reasoning, transparency, auditability, accountability, and robustness criteria to XAI techniques.
Macas et al. [33]	Covers CNNs, RNNs, LSTMs, GANs, and DRL across intrusion, malware, and botnets; highlights issues in explainability, robustness, and privacy.

The AI models examined in this study are drawn from a bounded set of high-coverage cybersecurity surveys and reviews (Table 3). Selection focused on peer-reviewed surveys published in well-established venues and widely cited within the cybersecurity community, with broad coverage of AI model families and explicit discussion of assurance-relevant properties such as explainability, traceability, robustness, accountability, or governance. Candidate studies were identified through major digital libraries (IEEE Xplore, ACM Digital Library, Scopus, and Google Scholar), and overlapping surveys were consolidated by retaining the most comprehensive and recent sources. The selected surveys serve to stabilise terminology and identify AI model families applied in cybersecurity. These models are organised into a taxonomy (Fig. 1) structured along two organising principles:

1. **Technical lineage:** which orders methods hierarchically from learning paradigms (e.g., ML, RL) to canonical model families and architectures (e.g., CNNs, GANs, LSTMs); and
2. **Operational scope:** which groups methods according to the cybersecurity operations in which they are applied, including language processing, knowledge representation and reasoning, and agentic or autonomous control.

The reference model, the AI taxonomy, and the AI candidate requirements provide the foundation for the methodology that follows, which derives authority criteria for AI systems in cybersecurity compliance and auditing.

4. Methodology

This section sets out the methodological procedure used to derive a set of evaluative criteria from binding cybersecurity compliance provisions. The study adopts a doctrinal legal research (DRL) approach, following the analytical steps described by Bhat [26], which treats authoritative regulations and standards as the primary objects of analysis and applies structured interpretation and synthesis to derive evaluative propositions.

The methodology consists of clause-level analysis of binding normative provisions drawn from a fixed corpus (First Layer Compliance) of cybersecurity regulations and standards. Each provision is interpreted using predefined decision anchors and mapped to the compliance–audit chain of *requirement* → *control* → *rule* → *evidence*, in order to determine the obligations imposed by the text and their implications for compliance determinations and evidentiary sufficiency. These doctrinal interpretations are then mapped with candidate Second-Layer AI requirements to identify recurring conditions that function as prerequisites for authoritative compliance and audit outcomes.

The approach operates over a fixed corpus and does not pursue statistical inference or population generalisation. Rigour rests on explicit decision anchors, consistent rule application, and full traceability

from clause text to classification and criterion derivation [26]. The methodology proceeds through the following steps.

4.1. Corpus delimitation

We restrict the corpus to binding provisions expressed in mandatory language (“shall”, “must”) within the regulations and standards used to construct the reference model (PCI DSS v4.0.1, DORA, UK GDPR, NIS2, ISO/IEC 27001:2022, and NIST SP 800-53 Rev. 5). Clauses framed as guidance, recitals, examples, or aspirational statements (“may”, “should”) are excluded because they lack enforceable force. Where a requirement is stated without an associated control clause, the requirement is retained as an operative anchor; such anchors are coded as Requirement and mapped to the reference-model operation(s) they constrain. We exclude provisions outside cybersecurity compliance and audit evidence conditions (e.g., general organisational governance not tied to control/evidence duties, and privacy provisions unrelated to security control obligations); where privacy provisions impose security or accountability duties that condition evidential admissibility (e.g., integrity/confidentiality and security of processing), they are retained. Table 4 reports retained provisions per standard/regulation after applying these rules.

4.2. Coding and interpretation

Each retained clause/sub-clause is analysed against the chain of *requirement* → *control* → *rule* → *evidence*. This provides an operational frame for interpreting the effect of provisions. Clauses and sub-clauses are coded according to their position in the chain, following the rules in Table 5.

All authors performed the classification using a shared codebook. Where interpretations diverged, the team resolved the issue by checking the clause text against the stated decision anchors and recorded the final decision in a decision log reflected in a dataset [35].

4.3. Derivation of criteria

We derived the CAAB criteria by starting from candidate AI-relevant requirements identified in the Second-Layer analysis (traceability, explainability, adaptability, and auditability) and testing them against recurring binding structures in the First-Layer corpus. A criterion is included in CAAB only where enforceable duties recur across the corpus and directly condition whether outputs qualify as admissible compliance artefacts or audit evidence within the *requirement* → *control* → *rule* → *evidence* chain. Where Second-Layer categories conflated distinct enforceable duties or operated at a misaligned level of abstraction, the candidate set was refined to reflect the enforceable structures expressed in the corpus.

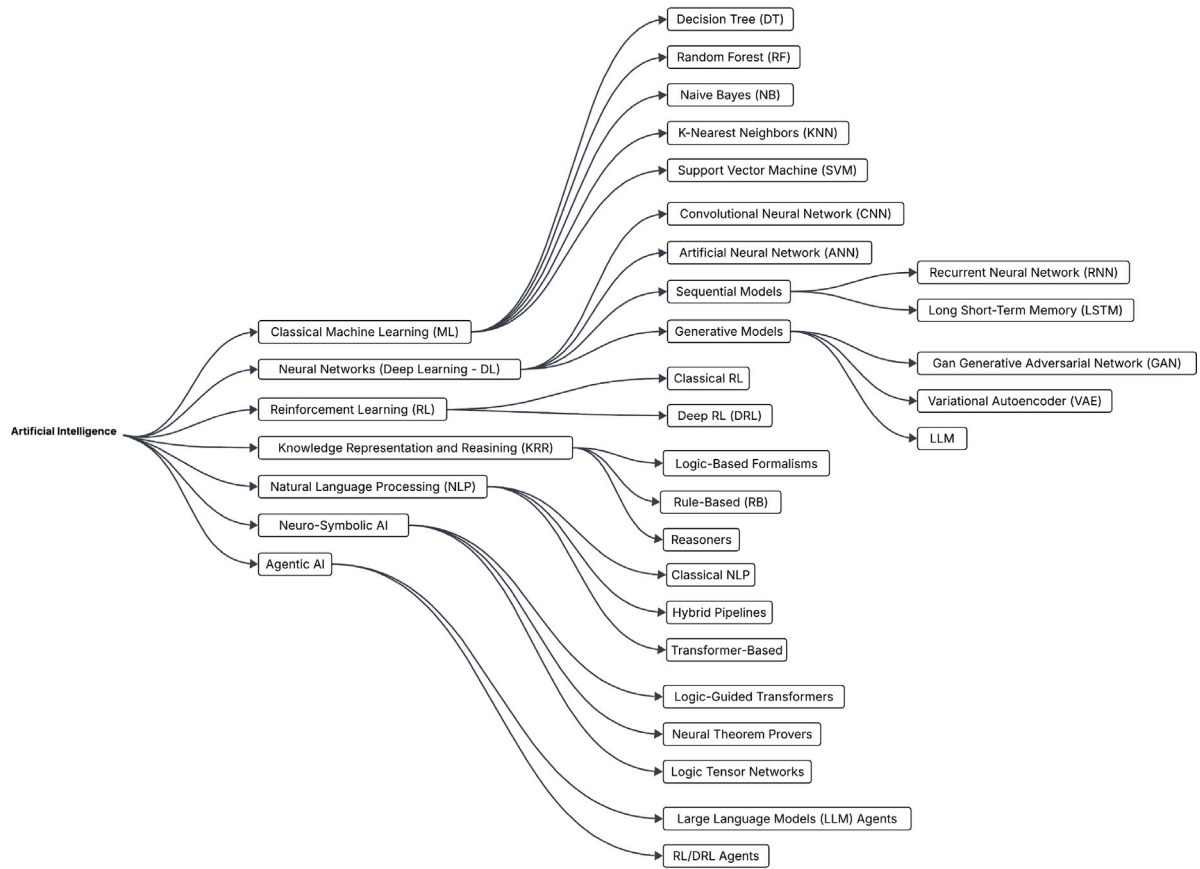


Fig. 1. Taxonomy of Artificial Intelligence approaches applied in Cybersecurity.

Table 4
Binding extraction scope across regulation and standards.

Framework	Binding extraction scope	Number of provisions extracted
PCI DSS v4.0.1	All 12 Requirements and sub-requirement	249
DORA (EU Reg. 2022/2554)	Chapter II — ICT risk management (Arts. 6–13) Chapter III — ICT incident reporting (Arts. 18–20) Chapter IV — Digital operational resilience testing (Arts. 25–27) Chapter V — ICT third-party risk management (Arts. 28–31/38–40) Chapter VI — Information-sharing arrangements (Art. 45)	144
UK GDPR	Article 5 — Principles (integrity & confidentiality) Article 25 — Data protection by design and default Article 32 — Security of processing Articles 33–34 — Breach notification	22
NIS2 (Directive 2022/2555)	Article 21 — Cybersecurity risk-management measures Article 23 — Reporting obligations Articles 28 — Database of domain name registration data Article 29 — Cybersecurity information-sharing arrangements Article 31–33 — Supervision and enforcement	56
ISO/IEC 27001	ISO/IEC 27001:2022 — ISMS requirements	98
NIST SP 800 Series	SP 800-53 Rev. 5 — Security and privacy controls	264

Within the corpus, *auditability* is primarily operationalised through duties governing artefact preservation, protection against modification, and integrity monitoring, and is therefore represented as *Evidence Integrity*. The corpus further distinguishes two enforceable patterns. Some clauses require determinations to be *derived* through explicit analytic evaluation, such as assessing defined factors or comparing results to criteria; these duties are coded as *Reasoning*. Other clauses require determinations to be *made intelligible* to reviewers by specifying rationale, scope, assumptions, methods, or report content; these duties are coded as *Explainability*. *Reasoning* constrains how conclusions are reached, whereas *Explainability* constrains how conclusions and their basis are

recorded for review. A clause may trigger both where it mandates analytic evaluation and its documentation. For example, NIST SP 800-53 Rev. 5 RA-3 (“Risk Assessment”) requires evaluation against defined factors, while RA-5 (“Vulnerability Monitoring and Scanning”) specifies the form and content of reviewable outputs.

Across the corpus, the validity of actions and their resulting records is repeatedly conditioned on authority and accountability structures, including defined roles, approvals or sign-off, segregation or escalation paths, and associated artefacts such as audit programme assignments or risk acceptance records. We represent these enforceable conditions as *Action Governance*.

Table 5
Interpretive chain for compliance clauses.

Position in chain	Rule of interpretation	Indicators considered	Exclusions
Requirement	Clause/sub-clause imposes a binding obligation stated in mandatory language.	“shall”, “must”; duties such as establish, implement, maintain.	“may”, “should”; guidance, recitals, or aspirational language.
Control	Clause/sub-clause specifies a measure or mechanism to realise a requirement.	Technical/procedural actions: configure, encrypt, restrict, monitor, review, train.	Abstract obligations with no implementable act.
Rule	Clause/sub-clause operationalises a control through explicit parameters.	Roles, approvals, thresholds, timings, workflows.	General control statements without operational detail.
Evidence	Clause/sub-clause requires artefacts that demonstrate fulfilment of a requirement, control, or rule.	Logs, records, reports, attestations, retention, audit trails.	Informational or descriptive clauses with no evidential consequence.

Table 6
Summary of the doctrinal derivation methodology.

Criterion	Working definition	Clause decision anchor	Indicative wording	Minimal evidential artefacts	Example fragments
Traceability	Condition to create a followable chain from a requirement clause/sub-clause through recorded determinations and control executions to evidential artefacts demonstrating fulfilment	Specified: binding verb + mapping/logging/retention. Indirect: outcome requires chain but clause lacks operative detail. Not specified: fulfilment without a chain possible.	Identify, record, retain, map, link, log, audit trail, SoA, chain of custody, register	Control–requirement matrix, decision register, SoA, unique IDs linking evidence to actions, audit logs	“shall retain audit logs...”, “shall map requirements to controls...”, “documented information shall demonstrate...”
Explainability	Condition to render outcomes understandable by stating criteria, scope, assumptions, method, rationale, or required report content	Specified: requires criteria, scope, justification, method, or report. Indirect: outcome presupposes reasons but none compelled. Not specified: result with no rationale or descriptive content.	Define criteria, scope, rationale, describe methods, document assumptions, justification, explainable	Written criteria/scope, rationale section, assumptions log, acceptance thresholds	“shall define criteria and scope...”, “report shall include reasons...”, “document assumptions used...”
Evidence Integrity	Condition to ensure artefacts remain authentic and detectably changed if modified	Specified: preservation, tamper-evidence, restricted modification, authentication, integrity monitoring. Indirect: retention required but no protection. Not specified: keep results without tamper protection.	Protect, preserve, tamper-evident, restrict modification, authenticate, verify, FIM, immutability	Write-protected stores, hashes with verification logs, FIM alerts, signatures, attestations	“shall protect logs from modification...”, “shall implement integrity monitoring...”, “shall preserve evidence...”
Adaptability	Condition to revise records, controls, or processes according to cadence or triggers (time, change, incidents)	Specified: requires review, update, change management, continual improvement with cadence/triggers. Indirect: change implied but cadence/triggers absent. Not specified: static one-off obligation.	Review, update, keep current, maintain, change management, continual improvement, revise upon change	Review calendar, change records with approvals, versioned policies, post-incident updates	“shall review at least annually...”, “shall update upon significant change...”, “maintained and kept current...”
Action Governance	Condition to allocate authority and define how, by whom, and when actions are taken and recorded	Specified: roles, approvals, segregation, escalation, acceptance, programme rules. Indirect: action requires approval/ownership but not stated. Not specified: instruction to act without role, approver, or timing.	Approve, authorise, assign responsibility, roles, escalation, prioritise, who/when	RACI for controls, approvals with timestamps, risk acceptance records, audit programme with methods	“obtain management approval...”, “roles and responsibilities shall be defined...”, “audit programme shall specify frequency...”
Reasoning	Condition to follow a declared analytic pathway that turns inputs into conclusions	Specified: analysis, evaluation against criteria, determination based on factors, prioritisation rules. Indirect: decision presupposes analysis but steps absent. Not specified: result demanded without analysis.	Analyse, assess, evaluate, compare with criteria, determine factors, prioritise rules, risk scoring	Documented analysis steps, factor weights, comparisons to thresholds, prioritisation ledger	“analyse consequences and likelihood...”, “evaluate by comparing results to criteria...”, “determine risk level from...”

From this induction process, six criteria are derived: Traceability, Explainability, Evidence Integrity, Adaptability, Action Governance, and Reasoning. Collectively, they specify the evidentiary and procedural conditions under which AI-mediated compliance and audit operations may produce outputs eligible for use in compliance determination and audit review.

Each criterion is coded as *Specified*, *Indirect*, or *Not specified* across First-Layer provisions, capturing the degree of normative instantiation. Provisions are also mapped to the compliance–audit reference model to support clause-level traceability and analysis of normative emphasis across operational processes.

Table 6 summarises the criteria, decision anchors, and indicative wording drawn from the corpus. Together, these constitute the Compliance–Audit Authority Benchmark (CAAB).

4.4. Demonstrator

The aim of CAAB is to serve as an evaluative benchmark for assessing AI systems in cybersecurity compliance and auditing. The demonstrator tests this role by applying CAAB to representative AI approaches and architectures, examining the extent to which they satisfy the criteria and identifying systematic gaps. The purpose is to show how CAAB operates as an evaluative lens and to indicate whether coverage of each criterion is supported, partially supported, or not supported within current practice.

4.4.1. CAAB × AI taxonomy (families)

AI approaches, organised by families, are first examined in their intrinsic design form against CAAB criteria. This baseline analysis considers whether the design characteristics of each family inherently support the criteria, without relying on auxiliary mechanisms such as post-hoc explainers, wrapper functions, or external logging frameworks.

4.4.2. CAAB × reference model

Representative AI deployments from the literature are situated within the operational processes of the reference model (e.g. RA, anomaly detection, evidence management). Within each operational process, deployments are evaluated against the CAAB criteria. This two-step mapping identifies both the operational locus of each deployment in the reference model and the degree to which the CAAB criteria are supported.

The literature mapping follows PRISMA-ScR as a reporting guide for a scoping review:

- **Keyword Strategy:** The search strategy was derived from the compliance reference model and refined iteratively to balance breadth and precision. Initial broad formulations (e.g., ("artificial intelligence" OR "machine learning") AND ("cybersecurity compliance and auditing")) produced high-volume, low-relevance results. To improve coverage and specificity, three axes were introduced:

1. **Operational functions:** terms reflecting compliance and audit domains such as vulnerability assessment, risk assessment, continuous monitoring, anomaly detection, control validation, and evidence management.

Example: ("reinforcement learning" OR "deep reinforcement learning") AND ("cybersecurity vulnerability assessment" OR "cybersecurity risk assessment") ("cybersecurity continuous monitoring" OR "cybersecurity anomaly detection" OR "cybersecurity attack detection" OR "malware detection") AND ("AI" OR "machine learning")

2. **AI model specificity:** targeted queries focusing on techniques linked to compliance-related controls.

Example: ("large language models" OR "transformers") AND ("policy" OR "rules" OR "controls")

3. **Regulatory anchoring:** integration of compliance frameworks and standards to capture system-level contributions in regulatory contexts.

Example: ("deep reinforcement learning" OR "Long Short-Term Memory") AND ("cybersecurity anomaly detection") AND ("PCI DSS" OR "ISO/IEC 27001" OR "NIST")

Where regulatory anchors produced no results, the anchor term was removed and the query re-run to allow broader retrieval.

Table 7
Inclusion and exclusion criteria.

Inclusion	Exclusion
Peer-reviewed journals and conferences; selected grey literature	Surveys, tutorials, position or vision papers
English language	Non-English
2018–2025	Pre-2018
Frameworks, models, or architectures operationalising compliance and audit functions	Pure prediction/detection without actuation or evidence path
Implemented systems or simulated prototypes showing enforcement, control actuation, policy decision, or evidence generation	Dataset-only, parameter-tuning, or conceptual-only works
Empirical evaluation (simulation, case study, deployment) or clear methodological/system design with direct applicability to compliance and audit functions	Duplicates and items failing screening

- **Inclusion and Exclusion Criteria:** To retain studies with direct operational relevance to the reference model, the following criteria were applied (Table 7).
- **Search and Selection:** Literature was identified through structured searches across major digital libraries (Google Scholar, Scopus, IEEE Xplore, ACM Digital Library, and SpringerLink). A purposive subset of studies was selected for analysis to represent system-level implementations that operationalise compliance and audit processes through automation (e.g., enforcement, control validation, evidence generation, and anomaly detection). Priority was given to peer-reviewed venues with established review standards.

5. Meta-compliance

This section applies CAAB to specify when AI systems may be regarded as authoritative in compliance and audit functions. The six criteria are grounded in doctrinal sources and operational relevance, and are used to evaluate representative AI approaches from the literature to assess how, and to what extent, these conditions are realised.

5.1. Compliance-Audit Authority Benchmark (CAAB)

CAAB is set to define the minimal conditions under which AI systems operating in control roles within cybersecurity compliance and auditing may be recognised as producing outputs with authoritative weight, understood as binding compliance artefacts capable of standing as audit evidence. The benchmark is not regulatory or legal. It specifies orthogonal conditions for recognising authority when AI systems execute or direct compliance and audit tasks without real-time human intervention; satisfaction of one criterion does not imply satisfaction of others. The benchmark focuses on whether outputs are authentic, interrogable, and reproducible.

Table 8 presents the criteria and their definitions. Each criterion is expressed in operational compliance and audit terms, treating AI systems as measures when deployed to execute or support these processes.

5.2. CAAB × AI taxonomy (families)

We evaluate AI families against the CAAB to determine whether authority is intrinsic to a model’s architecture or arises only through supplementary mechanisms. The scope also covers cryptographic mechanisms since, although not AI, they support integrity preservation and evidence verification across First-Layer compliance regimes [3–5]. Hash functions such as SHA-256 fix artefact digests and expose tamper

Table 8
Compliance–Audit Authority Benchmark (CAAB) criteria.

Criterion	Normative expectation
Traceability	Each implemented measure must have a verifiable chain linking it directly to its regulatory or policy requirement. The chain must establish provenance, scope, and justification with no ambiguity and must stand as evidence open to audit and scrutiny.
Explainability	Measures must produce determinations that a qualified human reviewer can fully understand without reliance on opaque or proprietary logic. Determinations must be stated with precision that provides full comprehension, supports independent verification, and remains subject to formal challenge in audit or oversight.
Evidence Integrity	Artefacts such as logs, scan outputs, change records, and attestations must be preserved in immutable and verifiable form, and must remain protected against tampering at every stage of creation, storage, and use. Artefacts that fail to demonstrate these criteria must not be recognised as valid evidence for compliance or audit.
Adaptability	Measures must remain responsive to regulatory amendments, evolving threat intelligence, and operational changes. The preservation of evidentiary integrity during adaptation is governed under the separate Evidence Integrity criterion.
Action Governance	Measures with delegated authority may autonomously select and execute actions. Such actions must remain bounded by rules or policies (symbolic, procedural, or contract-based). Measures must demonstrate that decisions follow these constraints and can be reviewed for accountability.
Reasoning	Measures must derive conclusions through a rigorous inference process that explicitly connects inputs to outcomes by applying mapped requirements, contextual facts, and structured logic. This process must ensure reproducibility under equivalent conditions, maintain full transparency of intermediate steps, and, where applicable, validate conclusions against alternative or counterfactual scenarios to confirm robustness.

events. Digital signatures bind artefacts to identities within PKI and assure authenticity and non-repudiation. Append-only ledgers and transparency logs maintain immutability through consensus and hash chaining. Smart contracts encode deterministic control rules within bounded scopes and form an auditable substrate for verifiable execution.

Within the CAAB criterion of Traceability, the literature separates algorithmic transparency from rule explicit traceability [102]. Algorithmic transparency refers to observability of model parameters (e.g., coefficients in linear models such as SVM) or posteriors (in probabilistic models such as NB), where these values are visible but do not reconstruct a decision path [40]. Rule explicit traceability refers to an auditable logical chain that connects inputs to outcomes, as in DT and RB systems where splits or clauses define explicit rules [103]. Neural networks expose numerical weights without semantic correspondence to inputs and provide transparency but not traceability in evidentiary terms [104]. We code T1 for algorithmic transparency and T2 for rule explicit traceability.

Explainability follows a comparable gradient. Symbolic and RB systems justify outputs through explicit rules or proofs and map to X1. Statistical and deep models depend on post hoc approximations such as SHAP, LIME or saliency mapping that estimate but do not reproduce internal reasoning and map to X2 [105]. Generative and language models produce fluent outputs and surface rationales that lack epistemic grounding [89] and map to X3.

Across all AI families, Evidence Integrity is non intrinsic. Models provide no built-in provenance, authenticity, or tamper protection. Integrity can be verified only at the artefact level, through external cryptographic measures such as hashing, signing, or ledgering, which ensure bit-level immutability but not the evidentiary integrity of the model's origin, training, or reasoning [106]. We classify this as EI1.

Adaptability describes how models revise behaviour when presented with new information. Incremental learners update from data streams [107] and map to AD1. RL and DRL adjust policies through feedback and reward optimisation [28] and map to AD2. Ontology and RB systems evolve through governed module or rule revision [59] and map to AD3.

Models demonstrate Action Governance when constraints act as first class elements of the decision calculus, such as integrity constraints, type or role restrictions or closed world policies, so that any derivation violating them is excluded by construction. In KRR approaches such as RB systems and verified reasoners, governance is intrinsic [31]. Explicit rules and ontological constraints delimit the search space and proof steps so that only policy compliant inferences or actions are produced and remain auditable.

Reasoning refers to explicit, reproducible inference chains, whether logical, probabilistic or hybrid [108,109]. Symbolic RL and neuro symbolic systems emit structured reasoning traces and map to R1. Sub symbolic learners such as CNN, RNN [40] and DRL produce results without interpretable inference and map to None [110]. Generative models, particularly LLMs, have introduced supplementary reasoning behaviour through techniques such as Chain of Thought (CoT) and Retrieval Augmented Generation (RAG) [95], which simulate inference sequences but do not establish formal logical or probabilistic chains. These methods remain post hoc and do not alter the intrinsic absence of explicit reasoning in such models.

Each capability is coded by type (T1, T2, X1, X2, X3, EI1, AD1, AD2, AD3, AG1, R1) and by support level (*Supported*, *Partially Supported*, *Not Supported*). The coding fixes a common baseline at the level of canonical implementations to ensure comparability across heterogeneous AI families. Specialised variants and auxiliary extensions are excluded to avoid conflating intrinsic architectural properties with properties supplied through external mechanisms.

Table 12 serves a structural role within the analysis. It distinguishes criteria that arise from a model family's internal decision calculus from criteria that must be supplied at the artefact level. In particular, it shows that Evidence Integrity is not intrinsic to AI model families and must be provided through external integrity controls, whilst rule-explicit traceability and explicit inference appear only where rules or constraints remain first-class elements of the model. This baseline helps support the interpretation of applied deployments in later sections and clarifies which authority conditions can be satisfied by design and which depend on additional evidentiary infrastructure.

Table 9 summarises the mapping.

5.3. CAAB × reference model

We evaluate the applied deployments retained after screening in Section 4.4.2. This set comprises 33 studies and is reported in Table 10, grouped by reference-model process. Each reviewed system is mapped to its corresponding reference-model process, and for every process-criterion pair we apply the same capability types used for AI families (T1–T2, X1–X3, EI1, AG1, R1, AD1–AD3) with support levels coded as *Supported*, *Partially Supported*, or *Not Supported*. A criterion is coded as *Supported* only when the study reports concrete mechanisms or artefacts that satisfy the relevant decision anchor in Table 6. Where such mechanisms or artefacts are not reported, the criterion is coded as *Not Supported*. *Partially Supported* is reserved for post hoc or auxiliary mechanisms that provide limited assurance scope. The tables are read column-wise to identify criteria that are systematically absent within each operational process, and row-wise to examine whether model-family properties carry through into operational deployments along the *requirement* → *control* → *rule* → *evidence* chain. Each study is also assessed using supplementary rubrics covering the AI operating model, dataset characteristics, and evaluation methodology.

Table 9
CAAB mapping across AI model families and integrity mechanisms (non-AI) (reason codes).

Model/Approach	T	X	EI	AD	AG	R
<i>Classical machine learning</i>						
SVM [36]	◦: T1	◦: X2	×	×	×	×
DT [37]	•: T2	•: X1	×	×	×	×
KNN [38,39]	◦: T1	◦: X2	×	◦: AD1	×	×
NB [40,41]	◦: T1	◦: X2	×	◦: AD1	×	×
RF [42]	◦: T1	◦: X2	×	×	×	×
<i>Reinforcement Learning</i>						
RL [28,43]	◦: T1	◦: X2	×	•: AD2	◦: AG1	◦: R1 (Policy)
DRL [44,45]	◦: T1	◦: X2	×	•: AD2	◦: AG1	×
<i>Neural networks (DL) [33,40,46–51]</i>						
LSTM	◦: T1	◦: X2	×	◦: AD1	×	×
RNN	◦: T1	◦: X2	×	◦: AD1	×	×
CNN	◦: T1	◦: X2	×	◦: AD1	×	×
ANN	◦: T1	◦: X2	×	◦: AD1	×	×
<i>Knowledge representation and reasoning [31,52–62]</i>						
Reasoners	•: T2	•: X1	×	◦: AD3	•: AG1	•: R1
Rule-Based	•: T2	•: X1	×	◦: AD3	•: AG1	•: R1
Logic Formalisms	•: T2	•: X1	×	◦: AD3	•: AG1	•: R1
<i>Natural language processing</i>						
Classical NLP [63,64]	•: T1/T2	•: X1	×	◦: AD1	×	◦: R1 (Rule)
Transformer Models [65–71]	◦: T1	◦: X2/X3	×	◦: AD1	×	×
<i>Neuro-symbolic AI [72–80]</i>						
Logic tensor nets [72,81]	◦: T1	◦: X1	×	◦: AD2	◦: AG1	•: R1
Neural theorem provers [82,83,83,84]	◦: T1	◦: X1	×	◦: AD2	◦: AG1	•: R1
Logic-guided transformers	◦: T1	◦: X1	×	◦: AD2	◦: AG1	•: R1
Hybrid pipelines	◦: T1	◦: X1	×	◦: AD2	◦: AG1	•: R1
<i>Generative models</i>						
GAN [85–88]	×	×	×	◦: AD1	×	×
LLM [27,68,89–93]	◦: T1	◦: X3	×	◦: AD1	×	×
<i>Agentic AI [94]</i>						
RL/DRL agents [28]	◦: T1	◦: X2	×	•: AD2	◦: AG1	◦: R1
LLM agents [95]	◦: T1	◦: X3	×	◦: AD3	◦: AG1	◦: R1
<i>Cryptographic mechanisms [96–99]</i>						
Hashing	•: T2	◦: X1	•: EI1	×	×	×
Signatures	•: T2	◦: X1	•: EI1	×	×	×
Blockchain/Ledger [100]	•: T2	◦: X1	•: EI1	◦: AD1	×	×
Smart Contracts [101]	•: T2	◦: X1	•: EI1	◦: AD1	•: AG1	×

Legend: • = Supported; intrinsic/inherent; ◦ = Partially Supported: achievable to some extent; × = Not Supported: absent.

Type codes: T1 = algorithmic transparency; T2 = rule-explicit traceability; X1 = symbolic; X2 = post hoc/wrapped; X3 = surface rationales; EI1 = cryptographic immutability/provenance; AD1 = incremental/online; AD2 = RL/DRL feedback; AD3 = ontology/KB modularity (governed); AG1 = embedded rules/contracts; R1 = explicit inference chain.

Cells are read as (support : capability type). If support is ×, no type code is shown. GANs remain × for T and X by default; X may be ◦ only in extended implementations (e.g., GAN Dissection).

6. Evaluation of findings

This section presents the evaluation, which synthesises two CAAB-based analyses: CAAB × AI Taxonomy and CAAB × Reference Model.

6.1. Evaluation of CAAB × reference model

Across operational processes, a shift from static artefacts to live workflows is evident; however, compliance authority depends on whether system outputs satisfy the CAAB conditions in Table 11. Table 8 defines CAAB as a normative benchmark, while Table 10 apply it to surveyed AI studies by mapping each study to reference-model functions and coding support for each criterion.

Table 11 synthesises this evaluation. For each reference-model process, it aggregates recurring CAAB gaps identified in Table 10 and specifies the minimal conditions under which process outputs qualify as audit-eligible artefacts within the requirement → control → rule → evidence chain. Each row corresponds to a reference-model process, including “Audit and Reporting”, and reports process-specific CAAB baselines.

Our assessment suggests that AI systems operating as implementers within First-Layer processes have gained operational depth, but they acquire authority only when designs satisfy the minimal conditions defined in CAAB. Where these conditions are absent, outputs may hold operational utility but remain non-authoritative in compliance terms.

6.2. Evaluation of CAAB × AI taxonomy

Regarding the inherent capacity of AI to satisfy the minimal CAAB baseline, our analysis demonstrates the patterns shown in Table 12:

The results show that Adaptability dominates the literature, particularly in relation to RL and DRL, while Reasoning traces and Action Governance are largely absent. Systems attain compliance authority only where, at minimum, symbolic governance, integrity services, and traceable justification are embedded at design stage.

6.3. Datasets and evaluation methods

Many studies (e.g., Arafah et al. [137]; Alam et al. [138]; Olawale and Ebadinezhad [136]) rely on legacy or simulated datasets such as

Table 10
CAAB mapping to reference model.

Study	Model	Dataset	Evaluation	T	X	EI	AG	R	AD
Vulnerability assessment									
Zolanvari et al. [111]	SVM, KNN, NB, RF, DT, ANN	SCADA/IoT testbed (451k, 80/20, multi-attack)	Accuracy, FAR, MCC, Confusion Matrix	o: T1	o: X2	×	×	×	×
Pham et al. [112]	DQN fuzzing (coverage-guided RL)	Binutils/Poppler (11 apps)	Edge coverage, unique paths, ER	o: T1	o: X1	×	o: AG1	o: R1	•: AD2
Muriithi et al. [113]	LSTM-DQN + iForest	HTV co-simulation plant	Torque, fuel, SoC metrics	o: T1	o: X2	×	o: AG1	o: R1	•: AD2
Lin et al. [114]	Bi-LSTM + Word2Vec + RF (cost-sensitive transfer)	SARD (83k vuln/52k clean) + OSS (FFmpeg, VLC, etc.)	Top-k Precision/Recall, t-SNE visualisation	o: T1	×	×	×	×	o: AD1
Zhang et al. [115]	GRU + CodeT5 + program slicing + vulDict + classifier	Big-Vul (188k funcs, 91 CWEs) + OSS validation	F1, Accuracy, Precision, Recall, Top-k localisation	o: T1	o: X2	×	×	×	o: AD1
Bahaa et al. [116]	GDistilBERT + 1D-CNN + BiLSTM	SARD subset (33k funcs, 7 CWEs) + OSS validation	Accuracy, F1, ROC, ablations	o: T1	×	×	×	×	o: AD1
Pandolfo et al. [117]	Ontology-based threat modelling (OWL-DL + SWRL + Pellet)	CAPEC, CWE, CVE, CPE (integrated KG)	GUI-backed reasoning; severity/impact rationale	o: T2	•: X1	×: EI1	•: AG1	•: R1	o: AD3
Sikder et al. [118]	LLaMA 3.2/StarCoderBase (1B-3B) + LoRA	ScrawID (9252 Solidity contracts) + Ince benchmark	Accuracy, Precision, Recall, F1, resource profiling	o: T1	o: X3	×	×	×	o: AD1
Risk assessment									
Radanliev et al. [119]	ANN, CNN, LSTM, Autoencoders, RL (cognitive edge engine)	Aggregated sources (Cisco Talos, PRC, SonicWall)	Monte Carlo simulation, FAIR-U correlation	o: T1	•: X1	×	o: AG1	•: R1	•: AD2
Park et al. [120]	Risk-Aware Problem Domain Ontology (PDO) with FAIR-based multi-perspective risk assessment	MITRE ATT&CK, Carbanak APT case study (15 asset-requirement items)	Expert scoring and aggregation per requirement/layer	o: T2	•: X1	×	o: AG1	o: R1	o: AD3
Vassilev et al. [121]	POMDP risk model (reduced to MDP); dynamic programming; ontology + policy model (OWL/SWRL/RDF)	No empirical dataset; parameterised scenarios with priors, transitions, costs, horizons	Analytical derivation with Bayes filtering, Bellman recursion, value function, decision tables	o: T1	•: X1	×	•: AG1	•: R1	o: AD3
Yao and García de Soto [122]	Conceptual ML risk assessment framework; ANN proposed; SHAP feature attribution; LLM Q&A support explored	Conceptual only; >20-paper review + 2 expert interviews; planned Monte Carlo synthetic data	No empirical evaluation; metrics listed (MAE, RMSE, Acc, F1; MSE/CE) but no implementation	o: T1	o: X2	×	×	o: R1	o: AD1
Requirement-to-control mapping and gap analysis									
Lee et al. [123]	ANN + Interpretive Structural Modelling (ISM) hybrid	Literature review; 65 survey responses; 17 expert panels; organisational case study	Empirical: ANN (RMSE, Pearson r, 10-fold CV); ISM via SSIM/Reachability/MICMAC; maturity appraisal	o: T1	o: X2	×	o: AG1	o: R1	o: AD1
Adebayo et al. [124]	Elasticsearch (Lucene) text search; CNN multilabel classification; Hybrid (ES+CNN) with SME-in-the-loop active learning	429 STIGs docs; 18,757 checks mapped to NIST 800-53 v4; text fields (title, description, rationale, fix); 3-fold split (15% test); +360 OpenShift rules (5 iterations)	Precision/Recall vs. thresholds; Hybrid recall > ES,CNN; CNN precision > Hybrid; F-score with SME feedback; CNN training timing	o: T1	o: X2	×	×	×	o: AD1
Elluri et al. [125]	OWL Protégé ontology with NLP (RAKE-nltk) for key-term extraction; Hermit reasoner	GDPR (Ch. 3-4); PCI DSS v3.2 (12 reqs); CSA Code of Conduct; validation corpus: AWS, Facebook, Google, Microsoft, WhatsApp	Three-phase process (keyword and modal extraction; GDPR/PCI → CSA mapping; ontology development; validation via provider policy instances, RDF); no quantitative metrics reported	o: T2	•: X1	×	•: AG1	•: R1	o: AD3

(continued on next page)

NSL-KDD, CICIDS 2017, UNSW-NB15, TON_IoT, and BoT-IoT, alongside laboratory metrics such as accuracy, precision, recall, and F1. These choices under-represent operational complexity and current regulatory

obligations. In numerous cases the data generation process, labelling policy, and anomaly definitions are not documented to an evidential standard, which weakens regulatory relevance.

Table 10 (continued).

Bar-Haim et al. [126]	RoBERTa classifier; SBERT (all-mpnet-base-v2) embeddings; score-sum ensemble for requirement → control text matching	14,188 reqs → 169 RCF categories (train/dev/test: 9929/2142/2076); subset: 2720 reqs → 280 NIST controls (6581/1411/1182 mappings); sources: AU, BR, CA, EU, FR, DE, JP, SG, UK, US; SME-curated over one year	Precision@1, Recall@10, MAP@10; error analysis; ablations on negatives, hierarchy, and domain pretraining	o: T1	o: X2	×	×	×	o: AD1
Barbara et al. [127]	Neuro-symbolic: Mask R-CNN (ResNet-101) + ASP (DLV2)	10k synthetic ECP images (320 × 320, augmentations); 32 real test images	Object detection (AP, AR, PR-AUC); symbolic reasoning (runtime, constraint satisfaction)	o: T1	o: X2	×	•: AG1	•: R1	o: AD3
Control implementation validation									
Static control validation via configuration & policy analysis: As systems that analyse control logic, system artefacts, and policy-rule associations to identify misconfigurations or violations before runtime.									
Jain et al. [128]	Node2Vec embeddings + RF, SVM, MLP	Synthetic SELinux policy corpora (125r/5c → 401r/10c → 455r/16c → final 10c)	Multi-class classification; ablations on class granularity; Acc, Prec, Rec, F1	o: T1	o: X2	×	×	o: R1	o: AD3
Alevizos [129]	SecureBERT + RF + DT; Hyperledger Fabric smart contracts	Lab sim (60 endpoints, transport/infrastructure); NIST CSF + ISO/IEC 27001 in chaincode; CTI (IBM X-Force, ATT&CK); ransomware/CVE scenario	Compliance Enforcement Rate (CER), Avg. Compliance Time (ACT); 95% CIs; t-tests vs. analyst baseline; variance analysis	•: T2	o: X2	•: EI1	•: AG1	o: R1	•: AD2
Dynamic Control Validation PT: As systems that validates the operational effectiveness of cybersecurity controls through adversarial behaviour									
Ghanem et al. [130]	Hierarchical RL over POMDP (IAPTF); cluster decomposition; Solve POMDP (GIP/PERSEUS)	VirtualBox corp LANs (2–200 hosts); 100 h: 25 clusters/102 vulns/80 exploits; 200 h: 52 clusters/153 vulns/115 exploits; prior belief reuse	Comparison vs. baseline RL and approximate solvers; policy-graph planning; Metasploit execution; reconfiguration/topology tests; PT benchmarking	o: T1	×	×	×	o: R1	•: AD2
Nguyen et al. [131]	RL (Q-learning/Replay) + DRL (DQN unstable) in PenGym vs. NASim/NASim(rev.)	PenGym range (CyRIS → NASim); scenarios: tiny/small-linear/medium-multi-site (max 16 hosts, 6 subnets, 192 actions); real Nmap/Metasploit; Linux-only, iptables firewalls	Train/test across NASim, NASim(rev.), PenGym; QL (4k eps), Replay (300 eps), DQN failed; 10 agents, 10 trials/scenario; metrics: success rate, avg. steps, exec/training time, stability (stdev); brute-force baseline	o: T1	o: X2	×	×	o: R1	•: AD2
Chowdhary et al. [132]	Conditional SeqGAN (GAN + RL policy gradient; semantic tokenisation + BPE)	XSS payload corpus (PayloadBox); BPE vocab; payloads relabelled via BurpSuite replays (ok/warn/fail/error)	Adversarial CGAN training (loss monitoring); BurpSuite replay validation; PT vs. ModSecurity (RB WAF) + AWS WAF	o: T1	×	×	×	×	o: AD2
Evidence Collection & Management									
Loumachi et al. [133]	Llama 3.1-8B (zero-shot) + RAG agent; mxbai-embed-large embeddings	Synthetic DFIR scenarios chunked into incident KB	DFIR metrics: accuracy, relevance, exact-match, top-k, human evaluation; successes and failures reported	o: T1	o: X2	×	×	o: R1	o: AD3
Pourvahab and Ekbatanifard [134]	SDN + Blockchain forensic architecture (DFeSB: SRVA auth, HSO keys, SA-DECC crypto, SHA-3 Merkle, FSC, LGoE)	Simulated IaaS (Java/CloudSim + ns-3.26); PoW miner; ≤100 users	Sim vs. CFLOG; metrics: response time, evidence insertion/verification, overhead, change rate; crypto timings (SHA-3 vs. (SHA-256) ²)	•: T2	o: X2	•: EI1	×	×	o: AD3

(continued on next page)

A further problem is temporal and regulatory drift. Datasets created under earlier frameworks (for example PCI DSS v2.0 or v3.0) embed classifications that no longer apply, producing models that inherit outdated violations as false positives while overlooking newly defined risks. The same drift affects compliance-derived artefacts such

as ontologies, OWL KB, and rule repositories. When these encode superseded versions, inferences remain internally consistent yet externally invalid with respect to present obligations. Operational datasets therefore risk freezing the threat landscape in time, whilst compliance-derived datasets risk freezing regulatory logic.

Table 10 (continued).

Brotsis et al. [135]	Blockchain for forensic evidence preservation (CTB, Hyperledger Fabric)	Permissioned Fabric + evDB; signed logs; hash-chained integrity proofs	Smart home/SOHO; off-chain evDB; on-chain metadata; chaincode ('CreateEvidence', 'TransferOwnership', 'EraseEvidence', 'GetEvidence'); ISP/LEA/ Prosecutor roles; chain-of-custody realisation; no benchmarking	o: T1	•: X1	•: E11	•: AG1	×	o: AD3
Continuous monitoring & anomaly detection									
Olawale and Ebadinezhad [136]	1D-CNN, LSTM, DT, RF, GB, SVM; IPFS + Ethereum (SHA-3/Keccak; AES/Fernet)	TON_IoT (35,976; 127 → 47f; 8c); Edge-IIoT (2.22M; 15c; 39f); UNSW-NB15 (2.54M; 206k/5k split; 44f)	80/20 split; SMOTE; grid search; Accuracy, Precision, Recall, F1; per-class reports	o: T1	o: X2	•: E11	×	×	o: AD1
Arafah et al. [137]	Enhanced BiGAN (dual encoders; shared generator/discriminator) + DL classifiers (RNN, GRU, LSTM)	NSL-KDD (official) and CICIDS-2017 (custom); balanced via synthetic E-BiGAN samples	Binary + multi-class evaluation: Accuracy, Precision, Recall, F1, ROC; synthetic data quality (similarity/discrimination); preprocessing (feature selection, encoding, scaling)	o: T1	×	×	×	×	o: AD1
Alam et al. [138]	DRL (DQN + stacked LSTM) with ϵ -greedy policy and experience replay	NF-BoT-IoT (600k), NF-ToN-IoT (1.37M), NF-UNSW-NB15 (1.62M), NF-UQ-NIDS (11.9M), NF-UNSW-NB15-v2 (2.39M); zero-day split excluding attack types	Zero-day training and multi-dataset validation; imbalance handled via KMeans-SMOTE vs. SMOTE/ADASYN; metrics: Accuracy, Precision, Recall, F1, Confusion Matrix; RL metrics (reward, convergence); LIME explanations	o: T1	o: X2	×	×	×	•: AD2
Incident Response (IR) & recovery									
Hammad et al. [139]	DQN, PPO, TD3 DRL agents (auto patch/block/isolate)	Aggregated "Awesome Cybersecurity Datasets"; 5 classes (Malware, Phishing, Intrusion, Advanced, Normal); 70/15/15 split; 5200 test samples	Accuracy, Precision, Recall, F1, Confusion Matrix	o: T1	×	×	×	×	•: AD2
Liu [140]	LLM-based multi-agent system (GPT-4o, AutoGen group chat + tool executor)	Backdoors & Breaches tabletop simulation (Core + Expansion); 6 team structures; 150 games; attack/procedure cards; inject/consultant excluded	Win rate by team structure, attack reveal rates, procedural effectiveness, ablations (team size, expertise mix)	o: T1	o: X3	×	•: AG1	o: R1	o: AD3
Castro et al. [141]	LLM blue agents (GPT-4o-mini, o1-mini, o3-mini, DeepSeek-V3) vs. RL baseline (KEEP: GNN-PPO); JSON comms protocol	CyBORG CAGE-4 simulation; FSM attacker variants (Aggressive, Stealthy, Impact, Degrade); 2 × 500-step episodes	Simulation-based evaluation in CAGE-4; LLM vs. RL; prompt variants (instruction, few-shot, role-based); post-hoc rationale clustering (PCA + KMeans)	o: T1	o: X3	×	o: AG1	o: R1	×
Audit and Reporting									
Gu et al. [142]	Foundation Model (GPT-4) with Chain-of-Thought prompting for auditing tasks	Artificial SAP FI/CO dataset (100 rows, 6 columns) for journal entry testing; additional financial statements and ACFR paragraphs for ratio analysis and extraction	Case study evaluation with reproducibility of prompt protocols; qualitative expert review of outputs; comparison against expected anomalies and calculations	o: T1	o: X3	×	o: AG1	•: R1	•: AD1
Chin et al. [143]	GPT-4 (LLM via ChatGPT, Chain-of-Thought prompting)	Synthetic SAP-style journal entries (100 rows, injected anomalies)	Qualitative anomaly detection accuracy, rationale inspection (No precision/recall/F1)	o: T1	o: X3	×	o: AG1	o: R1	•: AD2

Legend: • = Supported: intrinsic/inherent; o = Partially Supported: achievable to some extent; × = Not Supported: absent.

Type Codes: T1 = algorithmic transparency; T2 = rule-explicit traceability; X1 = symbolic; X2 = post hoc/wrapped; X3 = surface rationales; E11 = cryptographic immutability/provenance; AG1 = embedded rules/contracts; R1 = explicit inference chain; AD1 = incremental/online; AD2 = RL/DRL feedback; AD3 = ontology/KB modularity (governed).

Cells are read as (support : capability type). If support is ×, no type code is shown.

Evaluation practice is also narrow. Performance reporting is largely confined to laboratory indicators on artificial testbeds and binary outcomes, with limited attention to epistemic validity, provenance, and

reproducibility. In regulated settings, competence expectations parallel those in ISO/IEC 17025 [144], where environmental control, calibration, operator qualification, and process validation are explicit.

Table 11
Reference model operational processes and authority requirements.

Reference model function	Description	Authority requirements (minimal baseline)
Vulnerability assessment	Operates under incomplete signals and produces streams that change with system state.	Requires reproducible scans, provenance for configurations and signatures, and links from findings back to rules and requirements.
Risk assessment	Quantitative scoring improves clarity but creates dependency chains on prior states.	Depends on auditable model assumptions, versioned inputs, and reasoning traces that show how classifications lead to risk estimates.
Requirement-to-control mapping	Scale has improved, but unresolved ambiguity remains.	Requires semantics that ground correspondences, bidirectional links to clauses, and reviewable justification beyond statistical association.
Control implementation validation	Adaptive testing is operationally valuable but contingent on momentary conditions.	Needs policy-bound pass/fail criteria, reproducible test harnesses, and tamper-evident logs.
Evidence collection & Management	Durable artefacts exist alongside unverifiable generative outputs.	Requires integrity services for all derived artefacts, provenance for transformations, and clear separation between anchored records and summaries.
Continuous monitoring and anomaly detection	Behavioural modelling expands coverage yet often lacks confirmable events.	Needs explainable alert rationales, traceable features, and corroboration paths to independent evidence.
Incident response and recovery	Orchestration compresses time to action but can propagate errors.	Requires enforceable action policies, human-in-the-loop checkpoints for risky steps, and rollback with complete provenance.
Audit and reporting	Automation accelerates compilation, but certification remains human-driven.	Improves when reports preserve rule-level links, evidence integrity, and verifiable reasoning for key determinations.

Table 12
AI and authority requirements.

AI Cluster	Description	Authority Requirements (minimal baseline)
Symbolic (KRR, RB, Logic)	Provide native traceability, explainability, action governance, and reasoning, but adaptability remains limited and depends on disciplined maintenance.	Requires cryptographic integrity for artefacts, disciplined versioning, explicit mappings to requirements and controls, and provenance for inputs and inferences.
Classical ML	Effective for structured data and feature-based classification, but limited in reasoning, traceability, and governance.	Requires documented data lineage, reproducible pipelines, auditable parameter sets, and explicit links from features to compliance criteria.
Neural Network/DL	Offer strong adaptability and representation learning, but reasoning paths are opaque and governance is weak.	Needs governance policies, reproducible training and inference traces, provenance for artefacts, and justification beyond statistical correlation.
Transformers/Generative	Generalise across contexts and generate fluent outputs, but often unverifiable with weak traceability and integrity.	Requires versioned prompts, datasets, and models, reproducible inference contexts, provenance for generated artefacts, and safeguards against fabricated outputs.
RL/DRL	Adapt effectively under uncertainty, but behaviour depends on opaque reward signals and exploration history.	Requires explicit policy constraints, safety monitors, rollback with provenance, human-in-the-loop for risky actions, and tamper-evident logs.
Hybrid approaches	Combine symbolic governance with neural adaptability, achieving authority through the engineered pathway rather than the model family alone.	Requires end-to-end provenance, integrity services for all artefacts, enforceable policies at component interfaces, and auditable sources of retrieved or external evidence.

Evaluations that do not meet comparable standards risk overstating robustness. For authority-oriented assessment, performance metrics must be accompanied by evidence provenance, versioned inputs and models, auditable reasoning traces, and integrity guarantees for artefacts, in line with CAAB’s criteria.

6.4. Discussion

The CAAB-based evaluation confirms that no surveyed AI application achieves full authority across all criteria. Although the literature reports progress in accuracy, scalability, and efficiency, such advances remain confined to experimental contexts and disconnected from the evidential validity and regulatory assurances that define compliance legitimacy. Operational processes codified within standards as authoritative elements of compliance, including monitoring, detection, and

IR, are often reinterpreted as ordinary cybersecurity tasks. This reclassification strips them of their normative grounding and dissolves the conceptual foundation on which their authority should be assessed. However, under CAAB, such findings carry implications for both system design and governance:

6.4.1. Implications for system design

In regulated digital ecosystems, AI authority must be treated as an architectural property, not as an intrinsic attribute of individual models. Authority does not arise from statistical accuracy, but from system-level guarantees of lineage, integrity, traceability, and governed action. Components, whether symbolic or data-driven, cannot assert compliance authority in isolation. They acquire it only through integration with evidence, governance, and control services that render outputs reproducible and auditable across organisational boundaries.

From a system design perspective, this requires embedding compliance requirements directly into the operational fabric of the ecosystem. End-to-end decision and data lineage must be preserved across distributed infrastructures in a reviewable and tamper-resistant manner. Evidence services must cover inputs, transformations, models, policies, and outputs, while automated actions must operate within enforceable policy constraints, including approval, rollback, and segregation of duties. The result is a compliance-oriented architecture in which determinations are not only performant, but structurally sufficient for audit.

6.4.2. Implications for governance

Governance translates architectural sufficiency into enforceable obligations. An AI system can only be recognised as authoritative in audit when determinations are produced within an environment that preserves verifiable evidence, including replayability from preserved artefacts, cryptographically attested provenance, and a complete, attributable control history. These guarantees are not design preferences, but conditions for admissibility.

Regulators and auditors must therefore require auditor-facing interfaces that expose the artefacts, manifests, and verification material needed for independent inspection and re-execution. Retention rules must allocate custodianship and liability for compliance artefacts, specify durability and integrity guarantees, and define the consequences when replay or verification fails.

To ensure consistency across jurisdictions, governance frameworks should mandate conformance tests for replayability, chain-of-custody integrity, traceability to binding requirements, and enforceable revocation when dependent artefacts are invalidated. Risk-tiering can calibrate these obligations to impact. Authority in digital ecosystems is conferred through enforceable governance, not inferred from performance.

7. Systemic challenges, regulatory constraints, and research gaps

The analysis above defines the conditions for recognising AI authority in cybersecurity compliance and auditing. When these conditions are applied to current AI-based compliance and audit systems, recurring limitations emerge. This section identifies the constraints and research gaps exposed by evaluating existing practices against the CAAB criteria and the *requirement* → *control* → *rule* → *evidence* chain. The issues observed reflect structural misalignment between common AI design approaches and the requirements of authoritative compliance determination.

7.1. Challenges and constraints

7.1.1. Opaque systems in a rule-based domain

AI systems, particularly neural networks-based, used for cybersecurity compliance may detect anomalies in logs, traffic, or user behaviour, but their determinations lack policy awareness. Without explicit links from outputs to requirements, controls, and rules, reviewers cannot reconstruct why a condition is flagged as non-compliant. In domains defined by codified requirements, this opacity blocks traceability and reasoning, leaving outputs insufficient as audit evidence.

7.1.2. Control inference without control enforcement

Many AI cybersecurity compliance systems inspect configurations and policies, confirming that security artefacts are declared but not whether they are enforced in runtime. Compliance authority requires evidence of effective operation, scope coverage, monitoring, exception handling, and rollback. Systems that equate presence with enforcement present a distorted view of compliance and weaken evidential integrity.

7.1.3. From risk reduction to risk obfuscation

In compliance contexts, risk is defined by documented control deficiencies, regulatory exposure, and asset criticality, not by the absence of anomalies. AI compliance-audit systems that equate stable telemetry with low risk overlook missing or ineffective controls and create a false sense of assurance. Dashboards may indicate resilience whilst obligations remain unmet, leaving organisations exposed in audit.

7.1.4. Operational security by cosmetic compliance

AI-driven compliance automation frequently verifies encryption settings, ACLs, or logging declarations but stops short of validating their operational effectiveness. Authority depends on evidence that controls function under live conditions, across scope, with approval and rollback. Cosmetic indicators collapse in audit when effectiveness and coverage cannot be demonstrated.

7.1.5. Audit evidence as a trust proxy

Audit requires evidence that is traceable, reproducible, and explicitly tied to obligations. AI-generated alerts, classifications, or summaries rarely expose the provenance or reasoning needed for audit review. Without input lineage, rule references, and justification, outputs remain advisory rather than authoritative, and human oversight remains necessary to anchor compliance claims.

7.2. Identified gaps in current practices

7.2.1. Fragility of single-model AI dependence

When compliance-relevant judgements are produced by a single learned model or pipeline, drift, misclassification, or adversarial influence can directly shape determinations and associated evidence artefacts. This concentrates decision authority in a single opaque path. Within the CAAB framing, this highlights the importance of constrained judgement pathways and verifiable checks on outputs and actions.

7.2.2. Lack of provenance

AI generated artefacts rarely carry end-to-end provenance and integrity. Versioned inputs, model identifiers, inference traces, and action logs are seldom bound into a tamper-evident record. Without cryptographic protection and lifecycle linkage across creation, storage, and use, artefacts fail the minimal evidence conditions and cannot carry authority in compliance contexts.

7.2.3. Epistemic misalignment with regulatory logic

Cybersecurity compliance rests on explicit requirements and rule-governed reasoning. Many AI systems operate through statistical association that does not express determinations in rule-referenced terms. The result is internally plausible output that does not align with the logic auditors must review. Policy-aware inference and neuro-symbolic composition remain exceptions rather than the norm.

7.2.4. Undefined accountability in audit chains

Current regulations and standards anchor accountability in human actors but do not specify responsibility allocation when AI systems participate in compliance and audit decision-making. The literature likewise offers no settled framework as when outputs are inaccurate or incomplete, responsibility diffuses across developers, operators, and regulated entities, with no clear locus of fault. This diffusion conflicts with evidentiary and regulatory assurance chains, which require accountability to remain traceable, explicit, and legally assignable.

7.3. Answer to the research question

AI authority in cybersecurity compliance cannot be determined by technical criteria alone. Authority is an institutional status that depends on whether outputs produced by systems placed in control of compliance and audit actions are recognised as binding within assurance processes. Because compliance evidence follows a weakest-link logic [145], any defect along the *requirement* \rightarrow *control* \rightarrow *rule* \rightarrow *evidence* chain can negate authority even when model performance is high. Continuous assurance further constrains authority, as evolving threats and controls require preservation of evidential integrity and traceability.

We thus answer the research question by specifying a minimum-condition baseline through CAAB. CAAB defines the necessary pre-conditions that must hold before an AI system operating in control roles can be considered, in principle, for authoritative use in cybersecurity compliance or auditing. These conditions reflect the structural requirements of compliance determination itself, including traceability between outputs and controls, explainability for human review, integrity and provenance of evidence, controlled adaptability, governance of system actions and oversight, and reasoning that supports consistent and reviewable conclusions.

CAAB does not establish sufficiency or itself confer authority. Sufficiency remains context-dependent and institutionally determined and may be withdrawn in the presence of bright-line failures such as breaks in chain of custody, unlawful data use, or conflicts with binding rules. CAAB delineates the conditions that authority presupposes and provides a common evaluative basis for assessing AI systems deployed within cybersecurity compliance and audit workflows.

8. Conceptual solution, future direction and limitations

Meta-Compliance under CAAB makes authority contingent on verifiable reasoning, tamper-evident artefacts, enforceable action governance, and traceable links across the chain *requirement* \rightarrow *control* \rightarrow *rule* \rightarrow *evidence*. Symbolic methods natively support several of these properties; neural methods can meet them only when coupled with governance and integrity mechanisms that render outputs auditable. We therefore propose a Verifiable Reasoning Architecture (VRA) as a conceptual pathway.

VRA envisions AI systems that issue claims or execute actions only when cryptographically and logically verifiable against explicit evidence. Its purpose is to support contextual and semantic capability without sacrificing auditability. It is proposed as a research direction and minimal formal specification for AI operating in regulated environments.

8.1. Formal specification (minimal)

A VRA instance can be defined as a tuple:

$\langle K, E, P, R, \pi, X, V, \mathcal{L}, A \rangle$

where the minimal required components are:

- *K* (Knowledge schema): Ontologies and control vocabularies.
- *E* (Evidence store): Content-addressed blobs with cryptographic hashes and optional signatures.
- *P* (Perception modules): Neural extractors mapping raw inputs into symbolic forms, with digests.
- *R* (Rule base): Symbolic logic encoding domain constraints and policies.
- π (Planner/Decoder): Constrained generator producing candidate claims under the rules in *R*.
- *X* (Actioner): A semantic ML layer, combined with proof-based reasoning, responsible for autonomous actions (e.g., anomaly detection, evidence management, vulnerability assessment). Its behaviour is bounded by *R* and subject to *V*, so that the VRA not only reasons but also acts within sanctioned operational scopes.

- *V* (Verifier): Deterministic entailment and rule checking that evaluates the support of each claim relative to cited evidence.
- \mathcal{L} (Audit log): Chain-of-trust records of all inputs, intermediate states, and outputs.
- *A* (Attestation): Proof of the runtime integrity of the model and environment.

8.2. Axioms

The VRA class is defined by the following non-negotiable axioms:

1. **Evidence-closure:** No claim may be produced without citation to at least one evidence item in *E*.
2. **Deterministic inference:** Identical inputs, configurations, and evidence must yield identical outputs.
3. **Refutability:** Any claim failing verification in *V* or violating *R* is rejected or downgraded to “insufficient evidence”.
4. **Tamper-evidence:** All artefacts must be hashed and chained in \mathcal{L} .
5. **Attestability:** *A* must prove what code and weights actually executed.
6. **Least-authority generation:** The generative layer π cannot bypass *R* or *V*.
7. **Bounded autonomy:** Any action initiated by *X* must remain within constraints defined by *R* and must be subject to verification by *V*.

Component interaction. A VRA instance operates as a constrained evidence-to-determination loop. Raw inputs are transformed by *P* into structured representations linked to artefacts in *E* via the schema *K*. The planner π generates candidate determinations under rules *R*. The verifier *V* checks each candidate against (R, E, K) , rejecting or downgrading unsupported outputs. All inputs, states, and results are chained in the audit log \mathcal{L} . Where action is required, *X* executes only verified actions within *R* and records resulting artefacts back into *E*. Attestation *A* proves runtime integrity, enabling independent re-performance.

The axioms define the boundary conditions: every output must be evidence-backed, every reasoning step recorded, and all claims or actions reproducible and verifiable. The proposal remains conceptual; empirical instantiation and evaluation are future work.

8.3. Limitations

This study is theoretical and confined to doctrinal analysis. It presents a normative synthesis grounded in binding cybersecurity compliance requirements and does not claim to implement or empirically validate a system. As with doctrinal research, the findings depend on the interpretation and classification of authoritative texts within the selected corpus. Replicability therefore rests on the transparency of sources, extraction criteria, and coding rules, and reasonable disagreement over specific classifications or mappings remains possible [146]. The reference model captures high-level operational processes only and does not exhaustively represent the full depth of compliance practice, given the breadth and complexity of the domain. CAAB and the VRA are grounded in binding requirements but are advanced as baseline evaluative conditions and have not been validated or endorsed by regulators, auditors, or standard-setting bodies. Authority in practice also depends on factors outside the scope of this work, including system performance, operational reliability, organisational governance, and the specifics of assurance engagements. The corpus analysed is limited with other sectoral regimes and jurisdictional practices not examined. Standards evolve over time, which may introduce version drift. Future work will operationalise the benchmark into measurable indicators, test it in practice, and engage regulatory and policy makers to assess robustness and practical applicability.

9. Conclusion

This paper's doctrinal analysis shows that authority for AI in cybersecurity compliance is not a matter of statistical accuracy but of engineered guarantees: determinations must be traceable to sufficient evidence, reasoning must be reproducible, and actions must be explicitly governed. The CAAB distils these proof requirements into explicit criteria and, when applied first across AI families and then to AI systems operating in cybersecurity compliance and auditing processes, clarifies which architectures satisfy them intrinsically and which depend on external controls.

The analysis exposes a structural gap between First-Layer compliance, which addresses organisational requirements, and Second-Layer compliance, which addresses the conditions imposed on AI systems entrusted to enforce them. When these two layers converge, a Meta-Compliance regime arises in which not only the organisation but also the AI system, its training environment, datasets, networks, and operational context all become subjects of assurance. This recursive structure produces what may be described as a loop of inception: compliance generating further compliance, and assurance generating assurance of assurance. Without mechanisms that anchor each layer to sufficient evidence and auditable inference, the recursion risks obscuring the foundational aim of compliance, which is the safeguarding of data and systems.

The contribution of this study is therefore twofold. First, it provides an operational reference model for cybersecurity compliance and auditing evidentiary baseline against which both AI models and compliance processes can be assessed. Second, it proposes a conceptual VRA designed to ground AI outputs in cryptographically secured evidence, reproducible inference, and symbolic governance. In high-risk cybersecurity environments, compliance cannot be reduced to a checklist or equated with statistical accuracy. CAAB establishes the minimum evidentiary conditions for considering AI systems authoritative, but whether ultimate authority is recognised depends on how they operate under real oversight and how well their evidence holds when tested in practice.

CRedit authorship contribution statement

Fatma Yasmine Loumachi: Writing – original draft, Methodology, Investigation, Conceptualization. **Marcio J. Lacerda:** Writing – review & editing, Validation, Supervision, Methodology, Conceptualization. **Karim Ouazzane:** Writing – review & editing, Validation, Supervision, Methodology, Conceptualization. **Asma Adnane:** Writing – review & editing, Validation, Methodology, Conceptualization. **Ok-sana Adamyk:** Writing – review & editing, Validation, Methodology, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

References

- [1] D. Itani, R. Itani, A.A. Eltweri, A. Faccia, L. Wanganoo, Enhancing cybersecurity through compliance and auditing: A strategic approach to resilience, in: 2024 2nd International Conference on Cyber Resilience, ICCR, 2024, pp. 1–10.
- [2] J. Henriques, F. Caldeira, T. Cruz, P. Simões, A forensics and compliance auditing framework for critical infrastructure protection, *Int. J. Crit. Infrastruct. Prot.* 42 (2023) 100613.
- [3] PCI DSS v4.0.1, Payment card industry data security standard (PCI dss) v4.0.1, 2024.
- [4] ISO/IEC, Information technology — Security techniques — Information security management systems — requirements, 2022.
- [5] National Institute of Standards and Technology, NIST special publication 800-53 revision 5 update 1: Security and privacy controls for information systems and organizations, 2025.
- [6] International Organization for Standardization, ISO 19011:2018 - Guidelines for auditing management systems, 2018.
- [7] ISA, International standard on auditing 500: Audit evidence, 2009, Effective for audits of financial statements for periods beginning on or after December 15, 2004.
- [8] R. Amor, J. Dimyadi, The promise of automated compliance checking, *Dev. Built Environ.* 5 (2021) 100039.
- [9] B. Hutchinson, S. Dekker, A. Rae, Audit masquerade: How audits provide comfort rather than treatment for serious safety problems, *Saf. Sci.* 169 (2024) 106348.
- [10] European Parliament and Council of the European Union, Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation), 2016.
- [11] A. Marotta, S.E. Madnick, Analyzing the Interplay Between Regulatory Compliance and Cybersecurity, Working Paper, MIT Sloan School of Management, 2020.
- [12] N. Mohamed, Artificial intelligence and machine learning in cybersecurity: a deep dive into state-of-the-art techniques and future paradigms, *Knowl. Inf. Syst.* 67 (2025).
- [13] G. Deng, Y. Liu, V. Mayoral-Vilches, P. Liu, Y. Li, Y. Xu, T. Zhang, Y. Liu, M. Pinzger, S. Rass, PentestGPT: Evaluating and harnessing large language models for automated penetration testing, in: 33rd USENIX Security Symposium (USENIX Security 24), USENIX Association, Philadelphia, PA, USA, 2024, pp. 847–864.
- [14] C. Zhang, D. Song, Y. Chen, X. Feng, C. Lumezanu, W. Cheng, J. Ni, B. Zong, H. Chen, N.V. Chawla, A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data, in: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19), AAAI Press, 2019, pp. 1409–1416.
- [15] E. Papagiannidis, P. Mikalef, K. Conboy, Responsible artificial intelligence governance: A review and research framework, *J. Strateg. Inf. Syst.* 34 (2) (2025) 101885.
- [16] H.A. Haveman, R. Wetts, Organizational theory: From classical sociology to the 1970s, *Sociol. Compass* 13 (3) (2019) e12627.
- [17] O. of Management, Budget, M-25-21: Accelerating federal use of AI through innovation, governance, and public trust, 2025.
- [18] S.H. Cen, R. Alur, From transparency to accountability and back: A discussion of access and evidence in AI auditing, in: Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, EAAMO'24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 1–14.
- [19] High-Level Expert Group on Artificial Intelligence, Ethics Guidelines for Trustworthy AI, Technical Report, European Commission, Brussels, 2019, High-Level Expert Group on AI.
- [20] M.M. Saeed, M. Alsharidah, Security, privacy, and robustness for trustworthy AI systems: A review, *Comput. Electr. Eng.* 119 (2024) 109643.
- [21] G.L. Thanasas, G. Kampiotis, A. Karkantzou, Enhancing transparency and efficiency in auditing and regulatory compliance with disruptive technologies, *Theor. Econ. Lett.* 15 (2025) 214–233, Received: September 22, 2024; Accepted: February 7, 2025; Published: February 10, 2025.
- [22] W. Wang, S.M. Sadjadi, N. Rische, A survey of major cybersecurity compliance frameworks, in: 2024 IEEE 10th Conference on Big Data Security on Cloud (BigDataSecurity), 2024, pp. 23–34.
- [23] A. Marotta, S. Madnick, Convergence and divergence of regulatory compliance and cybersecurity, *Issues Inf. Syst.* 22 (1) (2021).
- [24] NIST Computer Security Resource Center, Anomaly, NIST Computer Security Resource Center (Glossary), 2025, Condition that deviates from expectations based on requirements specifications, design documents, user documents, or standards, or from someone's perceptions or experiences.
- [25] NIST Computer Security Resource Center, Behavioral Anomaly Detection, NIST Computer Security Resource Center (Glossary), 2025, Definition: A mechanism providing a multifaceted approach to detecting cybersecurity attacks.
- [26] P.I. Bhat, Doctrinal legal research as a means of synthesizing facts, thoughts, and legal principles, in: D. Watkins, M. Burton (Eds.), *Idea and Methods of Legal Research*, Oxford University Press, Oxford, 2020, pp. 143–168.
- [27] M.A. Ferrag, F. Alwahedi, A. Battah, B. Cherif, A. Mechri, N. Tihanyi, T. Bisztray, M. Debbah, Generative AI in cybersecurity: A comprehensive review of LLM applications and vulnerabilities, *Internet Things Cyber-Phys. Syst.* 5 (2025) 1–46.
- [28] H. Kheddar, Transformers and large language models for efficient intrusion detection systems: A comprehensive survey, *Inf. Fusion* 124 (2025) 103347.

- [29] H. Kheddar, D.W. Dawoud, A.I. Awad, Y. Himeur, M.K. Khan, Reinforcement-Learning-Based intrusion detection in communication networks: A review, *IEEE Commun. Surv. Tutor.* (2024) 1.
- [30] P. Manirihio, A.N. Mahmood, M.J.M. Chowdhury, A survey of recent advances in deep learning models for detecting malware in desktop and mobile platforms, *ACM Comput. Surv.* 56 (6) (2024).
- [31] L.F. Sikos, Cybersecurity knowledge graphs, *Knowl. Inf. Syst.* 65 (9) (2023) 3511–3531.
- [32] G. Rjoub, J. Bentahar, O. Abdel Wahab, R. Mizouni, A. Song, R. Cohen, H. Otrouk, A. Mourad, A survey on explainable artificial intelligence for cybersecurity, *IEEE Trans. Netw. Serv. Manag.* 20 (4) (2023) 5115–5140.
- [33] M. Macas, C. Wu, W. Fuertes, A survey on deep learning for cybersecurity: Progress, challenges, and opportunities, *Comput. Netw.* 212 (2022) 109032.
- [34] M. Brundage, S. Avin, J. Wang, H. Belfield, G. Krueger, G. Hadfield, H. Khlaaf, J. Yang, H. Toner, R. Fong, T. Maharaj, P.W. Koh, S. Hooker, J. Leung, A. Trask, E. Bluemke, J. Lebensold, C. O’Keefe, M. Koren, T. Ryyffel, J. Rubinovitz, T. Besiroglu, F. Carugati, J. Clark, P. Eckersley, S. de Haas, M. Johnson, B. Laurie, A. Ingerman, I. Krawczuk, A. Askill, R. Cammarota, A. Lohn, C. Stix, P. Henderson, L. Graham, C. Prunkl, B. Martin, E. Seger, N. Zilberman, S. Ó hÉigeartaigh, F. Kroeger, G. Sastry, R. Kagan, A. Weller, B. Tse, E. Barnes, A. Dafoe, P. Scharre, A. Herbert-Voss, M. Rasser, S. Sodhani, C. Flynn, T.K. Gilbert, L. Dyer, S. Khan, Y. Bengio, M. Anderljung, Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims, Technical Report, OpenAI, 2020.
- [35] AlinControl2025, AI in Control: Rethinking Cybersecurity Compliance and Auditing, GitHub repository, 2025, URL <https://github.com/AI-COMPLIANCE-AUDIT/AI-in-Control-Rethinking-Cybersecurity-Compliance-and-Auditing/tree/main>.
- [36] F. Shakerin, G. Gupta, White-box induction from SVM models: Explainable AI with logic programming, *Theory Pract. Log. Program.* 20 (5) (2020) 656–670.
- [37] A.B. Arrieta, N. Díaz-Rodríguez, J.D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, *Inf. Fusion* 58 (2020) 82–115.
- [38] O. Anava, K. Levy, K-Nearest neighbors: From global to local, in: D.D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Vol. 29, Curran Associates, Inc., 2016, pp. 4917–4925.
- [39] J. Haug, K. Broelemann, G. Kasneci, Dynamic model tree for interpretable data stream learning, in: *Proceedings of the 2022 IEEE 38th International Conference on Data Engineering, ICDE, IEEE, 2022*, pp. 2562–2574.
- [40] Z.C. Lipton, The mythos of model interpretability, *Commun. ACM* 61 (10) (2018) 36–43.
- [41] J. Gama, G. Castillo, Adaptive Bayes, in: F.J. Garijo, J.C. Riquelme, M. Toro (Eds.), *Advances in Artificial Intelligence — IBERAMIA 2002*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2002, pp. 765–774.
- [42] J. Rosaler, D. Desai, B. Sarmah, D. Vamvourellis, D. Onay, S. Pasquali, D. Mehta, Enhanced local explainability and trust scores with random forest proximities, in: *Proceedings of the 5th ACM International Conference on AI in Finance, ICAIF’24, 2024*, pp. 521–529.
- [43] M. Garnelo, K. Arulkumar, M. Shanahan, Towards deep symbolic reinforcement learning, 2016.
- [44] M. Landajuela, B.K. Petersen, S. Kim, C.P. Santiago, R. Glatt, T.N. Mundhenk, J.F. Pettit, D.M. Faissol, Discovering symbolic policies with deep reinforcement learning, in: *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, Vol. 139, PMLR, 2021, pp. 5979–5989.
- [45] J. Guo, J. Cheng, J. Cleland-Huang, Semantically enhanced software traceability using deep learning techniques, in: *2017 IEEE/ACM 39th International Conference on Software Engineering, ICSE, 2017*, pp. 3–14.
- [46] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, *Nature* 521 (7553) (2015) 436–444.
- [47] H. Turbé, M. Bjelogrić, C. Lovis, G. Mengaldo, Evaluation of Post-hoc interpretability methods in Time-Series classification, *Nat. Mach. Intell.* 5 (3) (2023) 250–260.
- [48] J. Zhang, W. Zhou, B.E. Ujcich, Provenance-Enabled explainable AI, *Proc. ACM Manag. Data* 2 (6) (2024).
- [49] Q. Zhang, L.T. Yang, Z. Chen, P. Li, A survey on deep learning for big data, *Inf. Fusion* 42 (2018) 146–157.
- [50] G. Marcus, Deep learning: A critical appraisal, 2018.
- [51] Y. Bengio, The consciousness prior, 2019.
- [52] J.S. Ribeiro, A. Nayak, R. Wassermann, Belief change and Non-Monotonic reasoning sans compactness, in: *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, AAAI Press, 2019, pp. 3019–3026.
- [53] H. Dibowski, Full traceability and provenance for knowledge graphs, in: *Formal Ontology in Information Systems (FOIS 2024)*, in: *Frontiers in Artificial Intelligence and Applications*, Vol. 394, IOS Press, 2024, pp. 223–237.
- [54] B.C. Grau, B. Motik, G. Stoilos, I. Horrocks, Completeness guarantees for incomplete ontology reasoners: theory and practice, *J. Artif. Int. Res.* 43 (1) (2012) 419–476.
- [55] A.M. Ileri, N. Rangarajan, J. Cannell, H. McGinty, VEL: A formally verified reasoner for OWL2 EL profile, 2024.
- [56] W.J. Clancey, The epistemology of a rule-based expert system — a framework for explanation, *Artificial Intelligence* 20 (3) (1983) 215–251.
- [57] W.R. Swartout, J.D. Moore, Explanation in second generation expert systems, in: J.-M. David, J.-P. Krivine, R. Simmons (Eds.), *Second Generation Expert Systems*, Springer Berlin Heidelberg, Berlin, Heidelberg, 1993, pp. 543–585.
- [58] J. McCarthy, P.J. Hayes, Some philosophical problems from the standpoint of artificial intelligence, in: B. Meltzer, D. Michie (Eds.), *Machine Intelligence*, Vol. 4, Edinburgh University Press, Edinburgh, 1969, pp. 463–502.
- [59] M. Peters, S. Sachweh, A. Zündorf, Large scale Rule-Based reasoning using a Laptop, in: F. Gandon, M. Sabou, H. Sack, C. d’Amato, R. Cudré-Mauroux, A. Zimmermann (Eds.), *The Semantic Web. Latest Advances and New Domains*, Springer International Publishing, Cham, 2015, pp. 104–118.
- [60] M. Richardson, P. Domingos, Markov logic networks, *Mach. Learn.* 62 (1) (2006) 107–136.
- [61] L. Getoor, Probabilistic soft logic: A scalable approach for Markov random fields over Continuous-Valued variables, in: L. Morgenstern, P. Stefaneas, F. cois Lévy, A. Wyner, A. Paschke (Eds.), *Theory, Practice, and Applications of Rules on the Web*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, p. 1.
- [62] V. Verreet, V. Derkinderen, P.Z.D. Martires, L.D. Raedt, Inference and learning with model uncertainty in probabilistic logic programs, in: *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI 2022)*, Vol. 36, Association for the Advancement of Artificial Intelligence (AAAI), 2022, pp. 10060–10069.
- [63] B. Teixeira, L. Carvalhais, T. Pinto, Z. Vale, Explainable AI framework for reliable and transparent automated energy management in buildings, *Energy Build.* 347 (2025) 116246.
- [64] M. Zhou, N. Duan, S. Liu, H.-Y. Shum, Progress in neural NLP: Modeling, learning, and reasoning, *Engineering* 6 (3) (2020) 275–290.
- [65] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K.Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems*, (NeurIPS 2013), Vol. 26, Curran Associates, Inc., 2013, pp. 3111–3119.
- [66] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186.
- [67] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are Few-Shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), *Advances in Neural Information Processing Systems*, Vol. 33, Curran Associates, Inc., 2020, pp. 1877–1901.
- [68] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, (NeurIPS 2017), Vol. 30, Curran Associates, Inc., 2017, pp. 5998–6008.
- [69] Y. Liu, H. Li, Y. Guo, C. Kong, J. Li, S. Wang, Rethinking Attention-Model explainability through faithfulness violation test, in: *Proceedings of the 39th International Conference on Machine Learning (ICML 2022)*, in: *Proceedings of Machine Learning Research*, Vol. 162, PMLR, 2022, pp. 7281–7292.
- [70] S.L. Blodgett, S. Barocas, H. Daumé III, H. Wallach, Language (technology) is power: A critical survey of “Bias” in NLP, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, 2020, pp. 5454–5476.
- [71] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, N.A. Smith, Don’t stop pretraining: Adapt language models to domains and tasks, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, 2020, pp. 8342–8360.
- [72] T. Rocktäschel, S. Riedel, End-to-end differentiable proving, in: I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, (NeurIPS 2017), Vol. 30, Curran Associates, Inc., 2017, pp. 3788–3800.
- [73] J. Xu, Z. Zhang, T. Friedman, Y. Liang, G. Van den Broeck, A semantic loss function for deep learning with symbolic knowledge, in: J. Dy, A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning*, in: *Proceedings of Machine Learning Research*, Vol. 80, PMLR, 2018, pp. 5502–5511.
- [74] Z. Hu, X. Ma, Z. Liu, E. Hovy, E. Xing, Harnessing deep neural networks with logic rules, 2020.

- [75] G. Spillo, C. Musto, M. de Gemmis, P. Lops, G. Semeraro, Recommender systems based on neuro-symbolic knowledge graph embeddings encoding first-order logic rules, *User Model. User-Adapt. Interact.* 34 (5) (2024) 2039–2083.
- [76] S. Bhardwaj, M. Dave, Integrating a Rule-Based approach to malware detection with an LSTM-Based feature selection technique, *SN Comput. Sci.* 4 (6) (2023) 737.
- [77] Z. Liu, Z. Wang, Y. Lin, H. Li, A Neural-Symbolic approach to natural language understanding, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 2159–2172.
- [78] P. Barbiero, G. Ciravegna, F. Giannini, M. Espinosa Zarlenga, L.C. Magister, A. Tonda, P. Lio, F. Precioso, M. Jamnik, G. Marra, Interpretable Neural-Symbolic concept reasoning, in: A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning*, in: *Proceedings of Machine Learning Research*, Vol. 202, PMLR, 2023, pp. 1801–1825.
- [79] X. Zhang, V.S. Sheng, Neuro-symbolic AI: Explainability, challenges, and future trends, 2024.
- [80] A. Baheri, C.O. Alm, Hierarchical Neuro-Symbolic decision transformer, 2025.
- [81] S. Badreddine, A. d'Avila Garcez, L. Serafini, M. Spranger, Logic tensor networks, *Artificial Intelligence* 303 (2022) 103649.
- [82] S. De Giorgis, A. Gangemi, A. Russo, Neurosymbolic graph enrichment for grounded world models, *Inf. Process. Manage.* 62 (4) (2025) 104127.
- [83] K. Yi, J. Wu, C. Gan, A. Torralba, P. Kohli, J.B. Tenenbaum, Neural-Symbolic VQA: Disentangling reasoning from vision and language understanding, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*, Curran Associates, Inc., 2018, pp. 1031–1042.
- [84] T. Eiter, N. Higuera, J. Oetsch, M. Pritz, A Neuro-Symbolic ASP pipeline for visual question answering, *Theory Pract. Log. Program.* 22 (5) (2022) 739–754.
- [85] R. Riaz, G. Han, K. Shaukat, N.U. Khan, H. Zhu, L. Wang, A novel ensemble wasserstein GAN framework for effective anomaly detection in industrial internet of things environments, *Sci. Rep.* 15 (1) (2025) 26786.
- [86] Z.C. Lipton, S. Tripathi, Precise recovery of latent vectors from generative adversarial networks, 2017.
- [87] L.J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, K.Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 27 (NeurIPS 2014)*, Curran Associates, Inc., 2014, pp. 2672–2680.
- [88] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, T. Aila, Training generative adversarial networks with limited data, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), *Advances in Neural Information Processing Systems*, Vol. 33, Curran Associates, Inc., 2020, pp. 12104–12114.
- [89] S. Wiegrefe, Y. Pinter, Attention is not not explanation, in: K. Inui, J. Jiang, V. Ng, X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 11–20.
- [90] K. Stechly, K. Valmeekam, S. Kambhampati, Chain of thoughtlessness? An analysis of CoT in planning, in: A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, C. Zhang (Eds.), *Advances in Neural Information Processing Systems*, Vol. 37, Curran Associates, Inc., 2024, pp. 29106–29141.
- [91] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, H. Wang, Retrieval-Augmented generation for large language models: A survey, 2024.
- [92] K. Mai, N. Ghate, J. Lee, R. Beuran, LLM-Based Fine-Grained ABAC policy generation, in: *Proceedings of the 11th International Conference on Information Systems Security and Privacy - Volume 2: ICISPP*, SciTePress, INSTICC, 2025, pp. 204–212.
- [93] C. Chen, D. Zhou, Y. Ye, T.J.-J. Li, Y. Yao, CLEAR: Towards contextual LLM-Empowered privacy policy analysis and risk generation for large language model applications, in: *Proceedings of the 30th International Conference on Intelligent User Interfaces, IUI'25*, Association for Computing Machinery, New York, NY, USA, 2025, pp. 277–297.
- [94] F. Piccialli, D. Chiaro, S. Sarwar, D. Cerciello, P. Qi, V. Mele, AgentAI: A comprehensive survey on autonomous agents in distributed AI for industry 4.0, *Expert Syst. Appl.* 291 (2025) 128404.
- [95] M.A. Ferrag, N. Tihanyi, M. Debbah, From LLM reasoning to autonomous AI agents: A comprehensive review, 2025.
- [96] N. Kumar, K. Kumar, A. Aeron, F. Verre, Blockchain technology in supply chain management: Innovations, applications, and challenges, *Telemat. Inform. Rep.* 18 (2025) 100204.
- [97] F. Bassan, M. Rabitti, From smart legal contracts to contracts on blockchain: An empirical investigation, *Comput. Law Secur. Rev.* 55 (2024) 106035.
- [98] S. Gorbunov, V. Vaikuntanathan, H. Wee, Predicate encryption for circuits from LWE, in: *Advances in Cryptology - CRYPTO 2015*, Springer-Verlag, Berlin, Heidelberg, 2022, pp. 503–523.
- [99] J. Wang, H. Wang, Monoxide: Scale out blockchains with asynchronous consensus zones, in: *16th USENIX Symposium on Networked Systems Design and Implementation, NSDI'19*, USENIX Association, Boston, MA, 2019, pp. 95–112.
- [100] A.K. Jain, N. Gupta, B.B. Gupta, A survey on scalable consensus algorithms for blockchain technology, *Cyber Secur. Appl.* 3 (2025) 100065.
- [101] M. Bartoletti, L. Galletta, M. Murgia, A minimal core calculus for solidity contracts, in: C. Pérez-Solà, G. Navarro-Arribas, A. Biryukov, J. Garcia-Alfaro (Eds.), *Data Privacy Management, Cryptocurrencies and Blockchain Technology*, Springer International Publishing, Cham, 2019, pp. 233–243.
- [102] J.A. Kroll, Outlining traceability: A principle for operationalizing accountability in computing systems, in: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FACT '21)*, Association for Computing Machinery, New York, NY, USA, 2021, pp. 758–771.
- [103] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining Black Box models, *ACM Comput. Surv.* 51 (5) (2018) 1–42.
- [104] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nat. Mach. Intell.* 1 (5) (2019) 206–215.
- [105] L.H. Gilpin, D. Bau, B.Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: An overview of interpretability of machine learning, in: *2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA, 2018*, pp. 80–89.
- [106] P.W. Grimm, M.R. Grossman, G.V. Cormack, Artificial intelligence as evidence, *Northwest J. Technol. Intelect. Prop.* 19 (1) (2021) 9–52.
- [107] J. Gama, I. Žliobaitundefined, A. Bifet, M. Pechenizkiy, A. Bouchachia, A survey on concept drift adaptation, *ACM Comput. Surv.* 46 (4) (2014) 1–37, <http://dx.doi.org/10.1145/2523813>.
- [108] National Institute of Standards and Technology (NIST), Artificial intelligence risk management framework (AI RMF) 1.0, 2023, Voluntary framework to manage AI risks across design, development, use, and evaluation; released January 26, 2023.
- [109] International Organization for Standardization, ISO/IEC 42001:2023 - information technology — Artificial intelligence — Management system, 2023.
- [110] Z. Cheng, J. Yu, X. Xing, A survey on explainable deep reinforcement learning, 2025.
- [111] M. Zolanvari, M.A. Teixeira, L. Gupta, K.M. Khan, R. Jain, Machine Learning-Based network vulnerability analysis of industrial internet of things, *IEEE Internet Things J.* 6 (4) (2019) 6822–6834.
- [112] V.-H. Pham, D. Thi Thu Hien, N. Phuc Chuong, P. Thanh Thai, P. The Duy, A Coverage-Guided fuzzing method for automatic software vulnerability detection using reinforcement Learning-Enabled Multi-Level input mutation, *IEEE Access* 12 (2024) 129064–129080.
- [113] G. Muriithi, B. Papari, A. Moghasssemi, A. Sundar, A. Arsalan, E. Buraimoh, L. Timilsina, G. Ozkan, C. Edrington, Vulnerability assessment and detection of stealthy sequential cyberattacks in hybrid tracked vehicles, *IEEE Trans. Transp. Electrif.* 11 (2) (2025) 6472–6489.
- [114] G. Lin, J. Zhang, W. Luo, L. Pan, O. De Vel, P. Montague, Y. Xiang, Software vulnerability discovery via learning Multi-Domain knowledge bases, *IEEE Trans. Dependable Secur. Comput.* 18 (5) (2021) 2469–2485.
- [115] X. Zhang, H. Guo, Z. Zhang, G. Tang, J. Sun, Y. Shen, J. Ma, Effectively detecting software vulnerabilities via leveraging features on program slices, *IEEE Internet Things J.* 12 (7) (2025) 8033–8048.
- [116] A. Bahaa, A.E.-R. Kamal, H. Fahmy, A.S. Ghoneim, DB-CBIL: A DistilBert-based transformer hybrid model using CNN and BiLSTM for software vulnerability detection, *IEEE Access* 12 (2024) 64446–64460.
- [117] L. Pandolfo, G. Corona, D. Guidotti, L. Pulina, A Knowledge-Driven approach to threat validation and security reasoning in modular systems, *IEEE Access* 13 (2025) 149817–149833.
- [118] F. Sikder, Y. Lei, Y. Ji, Efficient adaptation of large language models for smart contract vulnerability detection, in: *Proceedings of the 21st International Conference on Predictive Models and Data Analytics in Software Engineering, PROMISE'25*, Association for Computing Machinery, New York, NY, USA, 2025, pp. 65–74.
- [119] P. Radanliev, D. De Roure, K. Page, M. Van Kleek, O. Santos, L. Maddox, P. Burnap, E. Anthi, C. Maple, Design of a dynamic and self-adapting system, supported with artificial intelligence, machine learning and real-time intelligence for predictive cyber risk analytics in extreme environments 2013 cyber risk in the colonisation of Mars, *Saf. Extrem. Environ.* 2 (3) (2020) 219–230.
- [120] S.-H. Park, J.-W. Jung, S.-W. Lee, Multi-perspective APT attack risk assessment framework using Risk-Aware problem domain ontology, in: *2021 IEEE 29th International Requirements Engineering Conference Workshops, REW, 2021*, pp. 400–405.
- [121] V. Vassilev, D. Donchev, D. Tonchev, Risk assessment in transactions under threat as partially observable Markov decision process, in: *Optimization in Artificial Intelligence and Data Sciences*, in: AIRO Springer Series, Vol. 8, Springer, Cham, 2022, pp. 199–212.
- [122] D. Yao, B. García de Soto, Cyber risk assessment framework for the construction industry using machine learning techniques, *Buildings* 14 (6) (2024).

- [123] S.U. Lee, L. Dong, Z. Xing, M.E. Ahmed, S. Avgoustakis, Software security mapping framework: Operationalization of security requirements, 2025.
- [124] A. Adebayo, D. Sow, M.F. Bulut, Automated compliance blueprint optimization with artificial intelligence, 2022, arXiv:2206.11187.
- [125] L. Elluri, A. Nagar, K.P. Joshi, An integrated knowledge graph to automate GDPR and PCI DSS compliance, in: 2018 IEEE International Conference on Big Data (Big Data), 2018, pp. 1266–1271.
- [126] R. Bar-Haim, L. Eden, Y. Kantor, V. Agarwal, M. Devereux, N. Gupta, A. Kumar, M. Orbach, M. Zan, Towards automated assessment of organizational cybersecurity posture in cloud, in: Proceedings of the 6th Joint International Conference on Data Science & Management of Data (10th ACM IKDD CODS and 28th COMAD), in: CODS-COMAD '23, Association for Computing Machinery, New York, NY, USA, 2023, pp. 167–175.
- [127] V. Barbara, M. Guarascio, N. Leone, G. Manco, A. Quarta, F. Ricca, E. Ritacco, Neuro-Symbolic AI for compliance checking of electrical control panels, 2023.
- [128] K. Jain, P. Kapoor, J. Sum, A. Eaman, E. Hassan, E. Shakshuki, Using machine learning to analyze and detect anomalies in SELinux security policies, *Procedia Comput. Sci.* 257 (2025) 880–887, The 16th International Conference on Ambient Systems, Networks and Technologies Networks (ANT)/ the 8th International Conference on Emerging Data and Industry 4.0 (EDI40).
- [129] L. Alevizos, Automated cybersecurity compliance and threat response using AI, blockchain and smart contracts, *Int. J. Inf. Technol.* (2024).
- [130] M.C. Ghanem, T.M. Chen, E.G. Nepomuceno, Hierarchical reinforcement learning for efficient and effective automated penetration testing of large networks, *J. Intell. Inf. Syst.* (2022).
- [131] H.P.T. Nguyen, K. Hasegawa, K. Fukushima, R. Beuran, PenGym: Realistic training environment for reinforcement learning pentesting agents, *Comput. Secur.* 148 (2025) 104140.
- [132] A. Chowdhary, K. Jha, M. Zhao, Generative adversarial network (GAN)-Based autonomous penetration testing for web applications, *Sensors* 23 (18) (2023).
- [133] F.Y. Loumachi, M.C. Ghanem, M.A. Ferrag, Advancing cyber incident timeline analysis through Retrieval-Augmented generation and large language models, *Computers* 14 (2) (2025).
- [134] M. Pourvhab, G. Ekbatanfard, Digital forensics architecture for evidence collection and provenance preservation in IaaS cloud environment using SDN and blockchain technology, *IEEE Access* 7 (2019) 153349–153364.
- [135] S. Brotsis, N. Kolokotronis, K. Limniotis, S. Shiaeles, D. Kavallieros, E. Bellini, C. Pavu e, Blockchain solutions for forensic evidence preservation in IoT environments, in: 2019 IEEE Conference on Network Softwarization (NetSoft), 2019, pp. 110–114.
- [136] O.P. Olawale, S. Ebadinezhad, Cybersecurity anomaly detection: AI and ethereum blockchain for a secure and tamperproof IoHT data management, *IEEE Access* 12 (2024) 131605–131620.
- [137] M. Arafah, I. Phillips, A. Adnane, M. Alauthman, N. Aslam, An enhanced BiGAN architecture for network intrusion detection, *Knowl.-Based Syst.* 314 (2025) 113178.
- [138] K. Alam, M. Fahad Monir, M. Junayed Hossain, M. Shorif Uddin, M.T. Habib, Adaptive defense: Zero-Day attack detection in NIDS with deep reinforcement learning, *IEEE Access* 13 (2025) 116345–116361.
- [139] A.A. Hammad, S.R. Ahmed, M.K. Abdul-Hussein, M.R. Ahmed, D.A. Majeed, S. Algburi, Deep reinforcement learning for adaptive cyber defense in network security, in: Proceedings of the Cognitive Models and Artificial Intelligence Conference, AICCONF '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 292–297.
- [140] Z. Liu, AutoBnB: Multi-Agent incident response with large language models, in: 2025 13th International Symposium on Digital Forensics and Security, ISDFS, 2025, pp. 1–6.
- [141] S.R. Castro, R. Campbell, N. Lau, O. Villalobos, J. Duan, A.A. Cardenas, Large language models are autonomous cyber defenders, 2025.
- [142] H. Gu, M. Schreyer, K. Moffitt, M. Vasarhelyi, Artificial intelligence co-piloted auditing, *Int. J. Account. Inf. Syst.* 54 (2024) 100698.
- [143] J.H. Chin, P. Zhang, Y.X. Cheong, J. Pan, Automating security audit using large language model based agent: An exploration experiment, 2025.
- [144] ISO/IEC, ISO/IEC 17025:2017 — General requirements for the competence of testing and calibration laboratories, 2017, Confirmed in 2021.
- [145] C. Donalds, K.-M. Osei-Bryson, Cybersecurity compliance behavior: Exploring the influences of individual decision style and other antecedents, *Int. J. Inf. Manage.* 51 (2020) 102056.
- [146] S. Theil, Carefully tailored: Doctrinal methods and empirical contributions, *Oxf. J. Leg. Stud.* 45 (4) (2025) 1047–1075.