

OPEN ACCESS · DOUBLE-BLIND PEER REVIEW

International Journal of Security, Intelligence, Governance & Modern AI

IJSIGMA

VOLUME 1 · ISSUE 1 · 2026

Advancing Knowledge in Artificial Intelligence, Cyber Security, and Emerging Technologies

ISSN (Print): [ISSN to be assigned] · ISSN (Online): [ISSN to be assigned]

Published by Sigma Publisher · Islamabad, Pakistan · ijsigma.org

INTERNATIONAL JOURNAL OF SECURITY, INTELLIGENCE, GOVERNANCE & MODERN AI

(IJSIGMA)

Security · Intelligence · Governance · Modern AI

Advancing Knowledge in Artificial Intelligence, Cyber Security, and Emerging Technologies

Volume 1, Issue 1 (2026)

ISSN (Print): [ISSN to be assigned] · ISSN (Online): [ISSN to be assigned]
DOI (issue): [DOI to be assigned]

Islamabad, Pakistan

Copyright

© 2026 Sigma Publisher. All Rights Reserved.

All articles published in the International Journal of Security, Intelligence, Governance & Modern AI (IJSIGMA) can be quoted in future research with due acknowledgement, and the opinions expressed in published articles are those of the contributors.

Editorial Board

Editor-in-Chief

Prof. Dr. M. Mansoor Alam
Adjunct Professor, University of Technology Sydney, Australia
Specialization: Cyber Security

Associate Editors

Prof. Ghulam Ali Mallah
Postdoctoral Researcher, Glasgow
Specialization: Executive Director, IBCC

Professor Dato Dr. Mazliham Bin Mohd Suud
President, Multimedia University, Malaysia

Editorial Board Members

Prof. Dr. Abdul Razzaque MCS, NUST Islamabad (Postdoc UK (Cyber Security & AI))

Prof. Dr. Yasir Hafeez HOD, UIIT, PMAS Arid Agriculture University, Rawalpindi

Dr. Baber Majid Bhatti University of South Australia

Prof. Dr. Mushahid Anwer Martin-Luther-University Halle-Wittenberg

Prof. Dr. Adnan Khan Skyline University Dubai, UAE

Dr. Zunaira Jalil HOD, Air University Islamabad

Dr. Qaiser HOD, FAST University Islamabad

Asst. Prof. Dr. Sharreh Qazi University of the Punjab

Advisory Board

Prof. Ghulam Ali Mallah Postdoc Glasgow (Executive Director, IBCC)

Dr. Sadia Ali Assistant Professor (TTS), HEC Approved Supervisor (sadiaali@uaar.edu.pk)

Prof. Dr. Abdul Razzaque MCS, NUST Islamabad (Postdoc UK, Cyber Security & AI)

Dr. Baber Majid Bhatti University of South Australia

Publisher

Sigma Publisher
H1, Street 17, Block A
DHA Phase 1, Islamabad, Pakistan

Contact

Editorial Office: editor@ijsigma.org
General Inquiries: info@ijsigma.org

About IJSIGMA

Mission

To provide a premier platform for publishing high-quality, peer-reviewed research in Security, Intelligence, Governance, and Modern AI. We aim to foster innovation, accelerate knowledge exchange, and support the global research community with accessible, impactful studies.

Vision

To be recognized as a leading international journal in Security, Intelligence, Governance, and Modern AI, celebrated for academic excellence, open access principles, and meaningful contributions that shape research, policy, and technological innovation worldwide.

Scope

Security & Cyber Defense

Cybersecurity frameworks and advanced threat detection; cryptography, network, and cloud security; digital forensics and defense strategies; protection of critical infrastructure and digital assets; security policy development and governance models.

Intelligence & Information Analytics

Strategic, military, and open-source intelligence (OSINT); big data analytics for situational awareness and risk assessment; predictive modeling and decision-support systems; integration of intelligence systems for effective decision-making.

Governance, Policy & Digital Ethics

E-governance and digital transformation strategies; regulatory frameworks, compliance, and AI governance; privacy, data protection, and digital sovereignty; ethical, legal, and social implications of emerging technologies.

Modern AI & Computational Systems

Machine learning, deep learning, and generative AI applications; autonomous systems and intelligent agents; scalable computing architectures for governance and security; AI-driven automation and intelligent system design.

Open Access · Double-Blind Peer Review

All submissions undergo a thorough double-blind peer review process by experts in the field. Full open access ensures that research is freely available to readers worldwide.

Editorial

We are pleased to present Volume 1, Issue 1 of the International Journal of Security, Intelligence, Governance & Modern AI (IJSIGMA). This inaugural issue brings together five research papers that address pressing challenges in cyber security, AI literacy, misinformation, and the governance of autonomous AI systems.

Two papers focus on technical detection and evaluation integrity: a leakage-safe evaluation framework for gated autoencoder–Transformer multi-class network intrusion detection, and an intelligent adaptive threat-modeling approach for Advanced Persistent Threat detection in critical IoT networks. Together they emphasize transparent evaluation protocols and adaptive analysis for realistic, high-stakes environments.

The issue also addresses people, cognition, and institutions. One paper designs a scenario-based AI literacy toolkit for defence higher education; another examines AI-generated misinformation and the psychology of belief formation; and a fifth paper analyzes global governance of autonomous AI systems in cybercrime prevention.

Collectively, these studies show the value of combining technical rigour, human judgement, and institutional awareness in security and intelligent computing. We sincerely thank the authors for their valuable contributions, the reviewers for their insightful feedback, and the editorial team for their commitment to maintaining high academic standards. We invite scholars and practitioners to build on these insights and to contribute to future issues of IJSIGMA.

Prof. Dr. M. Mansoor Alam
Editor-in-Chief
International Journal of Security, Intelligence, Governance & Modern AI (IJSIGMA)

Disclaimer

The views expressed in these articles are solely those of the respective authors and do not necessarily reflect the views of the Editorial Board or the management and staff of Sigma Publisher. While every effort has been made to ensure the accuracy of the information provided by the authors, the Editorial Board does not accept any responsibility for any errors or omissions or breach of copyrights, if any.

Table of Contents

Volume 1, Issue 1 (2026)

Article Title / Author Names *Pages*

1. Designing a Scenario-Based AI Literacy Toolkit for Defense Higher Education.....01–33
Jawahir Che Mustapha, Nurul Akhmal Mohd Zulkefli and Mohd Nazri Ismail

2. A Leakage-Safe Evaluation Framework for Gated Autoencoder–Transformer Multi-Class Network Intrusion Detection 34–50
Azhar Javed, Naik Nawaz and Asher Mehfooz

3. AI-Generated Misinformation and the Psychology of Belief Formation: Investigating Cognitive Biases, Emotional Manipulation, and Susceptibility to Digital Radicalization 51–67
Saba Nadeem and Almas Irfan

4. Intelligent Adaptive Threat Modeling for Advanced Persistent Threat Detection in Critical IoT Networks 68–91
Salman Ghani Virk, Jawaid Iqbal1 and Tariq Hanif

5. Global Governance of Autonomous AI Systems in Cybercrime Prevention 92-108
Atif Ali and Zulqarnain Farid

ijsigma.org

SIGMA International Journal of Security, Intelligence, Governance & Modern AI

VOLUME 01 · ISSUE 01 · 2026 · ARTICLE 0001

RESEARCH ARTICLE OPEN ACCESS · ISSN: [ISSN to be assigned] · DOI: [DOI to be assigned]

Designing a Scenario-Based AI Literacy Toolkit for Defense Higher Education

Jawahir Che Mustapha1,* Nurul Akhmal Mohd Zulkefli2 and Mohd Nazri Ismail3

1 Malaysian Institute of Information Technology, Universiti Kuala Lumpur, 1016, Jalan Sultan Ismail, Kuala Lumpur, 50250, Malaysia

2 Dhofar University, Salalah, Oman

3 National Defence University of Malaysia

Corresponding author: Jawahir Che Mustapha (jawahir@unikl.edu.my) · ORCID: [0000-0000-0000-0000]

Author emails: Nurul Akhmal Mohd Zulkefli (nzulkefli@du.edu.om); Mohd Nazri Ismail (m.nazri@upnm.edu.my)

ABSTRACT

Artificial intelligence (AI) is increasingly affecting higher education and professional decisions, thus calling for a broader form of AI literacy that includes not only technological expertise but critical analysis, human agency, verification, calibrated reliance, and ethical responsibility. To address this need, this study aims to develop a Scenario-Based AI Literacy Toolkit for Defense Higher Education, especially taking into account the demands of high-stakes learning and Professional Military Education (PME). An evidence-informed, staged approach was selected in which preliminary evidence synthesis was combined with a prospective validation protocol. The preliminary search revealed 515 records, out of which 514 survived the deduplication process, and 204 became potentially eligible for further assessment; however, since no full-text eligibility process was completed, those records could not be considered a systematic evidence corpus. A verified subset of 20 evidence items was used to detect recurring constructs, assessment approaches, Defence/PME evidence, and methodological gaps. The synthesis revealed that AI literacy is multidimensional and responsible AI judgement requires more than confidence or skill proficiency. Hence, six tentative dimensions of the toolkit were developed: (1) understanding and limits of AI, (2) critical analysis and judgement, (3) human agency and intellectual ownership, (4) reliance calibration, (5) verification, and (6) ethical and responsible judgement. The synthesis also reveals certain evidence gaps, including prevalence of self-reports, lack of objective measures of scenario-based learning outcomes, fragmented assessment of verification and calibrated reliance, and limited evidence on direct Defence/PME validation. The toolkit is presented here as an evidence-informed design architecture rather than as a validated educational intervention or psychometric instrument. Further research will require validation from experts, cognitive interviews, piloting of scenarios, performance assessment, reliability and psychometric testing, experimental

validation, and longitudinal Defence/PME assessment.

Keywords: Artificial Intelligence Literacy; Scenario-Based Assessment; Defense Higher Education; Professional Military Education; Human AI Judgement; Human Agency;

Highlights

- **Artificial intelligence (AI) is increasingly affecting higher education and…**
- **To address this need, this study aims to develop a Scenario-Based AI Literacy…**
- **An evidence-informed, staged approach was selected in which preliminary evidence…**

IMPACT STATEMENT

Artificial intelligence (AI) is increasingly affecting higher education and professional decisions, thus calling for a broader form of AI literacy that includes not only technological expertise but critical analysis, human agency, verification, calibrated reliance, and ethical responsibility. To address this need, this study aims to develop a Scenario-Based AI Literacy Toolkit for Defense Higher Education, especially taking into account the demands of high-stakes learning and Professional Military Education (PME). An evidence-informed, staged approach was selected in which preliminary evidence synthesis was combined with a prospective validation protocol. The preliminary search revealed 515 records, out of which 514 survived the deduplication process, and 204 became potentially eligible for further assessment; however, since no full-text eligibility process was completed, those records could not be considered a systematic evidence corpus. A verified subset of 20 evidence items was used to detect recurring constructs, assessment approaches, Defence/PME evidence, and methodological gaps. The synthesis revealed…

I. INTRODUCTION

In recent years, artificial intelligence (AI) has been changing the way learners gain information, create content, analyze evidence, solve problems, and make decisions in higher education. The development of generative artificial intelligence (GenAI) has sped up this process since current AI systems can generate seemingly authoritative text, analysis, explanation, summary, and recommendations in just a few seconds [1]. Consequently, AI literacy is now not about using the technology properly; it is also about understanding AI's potential and limitations, evaluating AI-generated outputs, validating information against available evidence, identifying uncertainties and biases, and maintaining accountability for decisions made with the help of AI. Indeed, previous reviews of AI literacy show that it is multidimensional and includes several aspects, namely, knowledge, application, evaluation, ethics, metacognition, and context-awareness [2]. The significance of this shift is especially pronounced in Defense Higher Education, where the educational tasks often imply making professional judgments in a complicated information environment with a considerable amount of uncertainty and making decisions with potentially serious consequences. Under these circumstances, inappropriate use of an AI recommendation can have much more dramatic consequences compared to giving an incorrect answer during a normal educational activity [3,4]. Therefore, the goal of AI education in this context should not only be encouraging learners to use an AI tool effectively. It should also be developing skills that allow determining when AI support is needed, when an AI-generated output should be challenged, when some claims require validation, when output needs to be modified or rejected, and when human judgment should prevail. High-stakes and military research contexts have revealed the importance of automation bias, algorithm aversion, calibrated trust, and human supervision in decision-making processes supported by AI.

A key issue, however, is that current efforts to teach AI literacy may fail to assess whether students are able to convert their knowledge into correct actions. Existing research on AI literacy recognizes multiple aspects of competence; yet measurement often rests on the learners' perception of their own knowledge, confidence, or competence [6]. According to the findings of AI-literacy reviews, there is a notable difference between the learners' perceptions of what they are able to do and the real actions they are capable of performing. For the case of Defense Higher Education, such a difference takes special importance. The learners may know that the AI system produces wrong information, but still accept its claims because of their plausibility. Likewise, the learners may feel confident in using AI; at the same time, they will be unable to detect uncertainty or change their dependence in the case of high risks in particular tasks. As will be discussed further in this Introduction, such a situation represents three interrelated gaps which become the motivation for the present study: the gap in conceptualization of AI literacy and responsible judgement, the gap in measurement of the judgement itself, and the gap in the context of direct Defence/PME-related evidence.

The educational issue, in other words, is not whether the learner knows about AI, but whether he or she is capable of exercising responsible judgment in response to AI outputs. To be responsibly literate in AI means that the learner can move from knowing to doing, which includes understanding the capabilities and constraints of the AI system, evaluating the quality of the output generated by the system, verifying claims by checking them against authoritative sources, matching reliance on the AI system to its capabilities and the stakes involved, and being ultimately accountable for the resulting decision. There is ample evidence in favor of measuring these behaviors within decision-making sequences involving accept, question, verify, modify, override, or reject choices, rather than using trust or confidence as an adequate measure of AI competency. This makes the case for researching scenario-based AI literacy education. In the scenario-based method, one can put the learner into a situation where AI outputs are reliable, incomplete, uncertain, biased, or not relevant to the problem at hand. Rather than asking the learner to self-report on whether he or she is a competent user of AI systems, the scenario-based measure will allow one to assess the learner's reaction to an AI output that is either helpful, misleading, incomplete, contradictory, or uncertain.

The necessity for scenario-based assessments is especially important in the context of the transfer because Defense Higher Education is a high-risk transfer context and not merely another example of using higher education AI literacy in general. In the former case, general higher education literature can contribute to understanding of educational mechanisms, whereas military and professional literature can help understand the concepts of reliance, automation bias, challenge, and consequence management. However, the latter cannot automatically imply what defense students know and how they act. Transfer depends on factors such as learner demographics, task, authority, institutional policy, assessment circumstances, and consequence structure. The conceptual logic outlined by the paper suggests that AI literacy is a judgment process pathway and not a set of separate technical skills. The suggested conceptual framework is supposed to be a conceptual architecture and not necessarily a verified causative model. The above-mentioned distinction is consistent with the existing evidence on the relevance of agency, critical thinking, reliance, verification, and accountability, but not on their developmental order.

Table I presents a structured summary for this topic. It compares traditional awareness-based AI literacy to performance-based criteria, and explains why the toolkit should be more focused on reactions rather than confidence.

Table I. From Conventional AI Literacy to Scenario-Based AI Literacy:

Conventional AI-literacy emphasis	Key limitation	Scenario-based requirement
Understanding AI	Knowledge may not translate into behavior	Demonstrate appropriate AI use in context
Awareness of AI limitations	Learners may recognize limitations but ignore them	Identify limitations during an actual task
Confidence in AI use	Confidence does not demonstrate competence	Demonstrate observable performance
Trust in AI	Excessive trust can produce automation bias	Calibrate reliance according to evidence and stakes
Ethical awareness	Knowing principles does not guarantee ethical action	Apply ethical judgement in realistic situations
AI-generated information	Plausible outputs may remain inaccurate	Verify claims against authoritative sources
General AI skills	Context and consequences may be overlooked	Adapt AI use to task and consequence conditions
Self-report assessment	May overestimate actual competence	Combine performance, explanation, and verification

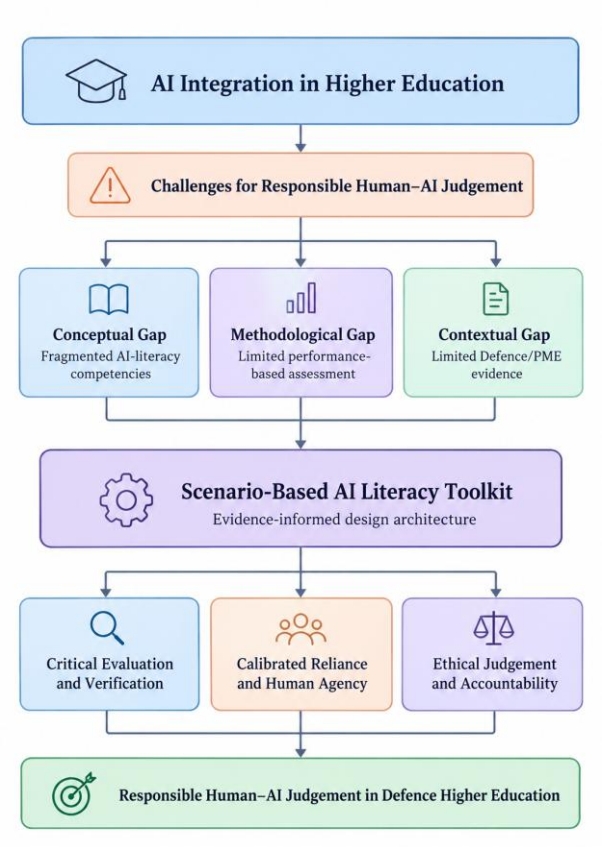


Figure 1. Conceptual overview of the rationale for scenario-based AI literacy in Defense Higher Education.

Figure 1 above depicts the gaps in the literature on AI literacy, which have spurred the current research, including fragmented competencies, lack of performance assessment, and a lack of specific evidence from Defence/PME. The figure illustrates the relationship between these gaps

and the proposed scenario-based toolkit, which highlights critical thinking, appropriate trust, human autonomy, and responsible judgment.

A. Research Gap

Firstly, there is a conceptual issue with AI literacy and responsible human judgement, as current frameworks consider various dimensions of AI literacy, but those dimensions do not always cohere as part of a decision-making process where learners assess AI output and respond accordingly. As a result, knowledge, skills, attitude, trust, ethical considerations, and critical thinking can remain separate competencies rather than elements of a joint decision-making process. Secondly, there is a methodological gap as far as assessment of AI literacy is concerned. Current data show that self-reporting measures are widely used, while objective or scenario-based assessment is underdeveloped. As a result, educational programs run a risk of measuring students' perceptions of competence instead of assessing their ability to verify the information, understand uncertainty, calibrate trust, and dispute incorrect AI suggestions. Thirdly, there is a gap as far as Defense Higher Education is concerned, as current defense and military literature considers the importance of responsible AI and human supervision and accountability, while higher education literature covers data on AI literacy and educational techniques. Yet there is no research into Defense learners and how they should be taught and tested in the context of responsible AI-based judgement.

B. Purpose of the Study

In the context of this study, the intention is to generate an evidence-informed Scenario-Based AI Literacy Toolkit for Defense Higher Education, which will take complex concepts related to AI literacy and articulate these into concrete scenarios, observable learner behavior, assessment criteria, and reflective decision-making exercises. In developing this toolkit, the assumption will be made that in high-stakes education, AI literacy should be reflected in the quality of human judgment used in regard to the outcomes of AI technology, and not merely inferred on the basis of self-report measures such as knowledge and confidence. Hence, this proposed toolkit should be considered as an evidence-informed, testable pedagogic design rather than a validated intervention.

C. Contributions of the Study

In the field of responsible AI education in Defense Higher Education, the present study makes a theoretical contribution to the existing body of knowledge by proposing an evidence-informed Scenario-Based AI Literacy Toolkit that links together the concepts of AI literacy, critical evaluation, evidence verification, calibrated reliance, human agency, and accountability. The study makes a methodological contribution by developing the Scenario-Based AI Literacy Toolkit and translating the abstract concept of AI literacy into observable human behavior that can be studied via realistic scenarios. The practical contribution made by the study lies in providing the architecture for designing curriculum, developing scenarios, performance assessments, faculty training, and learner self-reflection. The key contribution of the study is, thus, the integration and contextualization of scenario-based AI literacy in Defense Higher Education, rather than the novel nature of the dimensions of AI literacy per se. It should be stressed that the proposed educational architecture is presented as an evidence-informed and testable one, not a validated intervention, due to the necessity to validate the proposed intervention itself through expert validation, scenario piloting, psychometric testing, and Defence/PME-specific research.

II. LITERATURE REVIEW

The review integrates AI literacy frameworks, scale validation, interventions, human–automation interactions, and examples from the military context. These literature types answer different

questions and cannot be used interchangeably: reviews inform about competencies, scale validations validate scales, interventions explore learning, and military contexts highlight reliance in situations of great importance. For that reason, the synthesis starts by comparing the population, design, assessment type, and inference possibilities of each type of evidence before identifying the requirements for the suggested toolkit.

A. Conceptualizing AI Literacy in Higher Education

AI literacy has come to be regarded as a significant educational construct given that artificial intelligence systems have been increasingly incorporated into various fields of academia, profession, and society. When AI literacy was first framed, it was mainly concerned with the ability of the learners to understand concepts related to AI. However, these days, it has evolved into an interdisciplinary skill involving the knowledge, application, assessment, creation, and ethical interaction of learners. Long and Magerko presented an early competency-focused definition of AI literacy by highlighting the need for learners to be able to critically evaluate and interact with AI technologies [8]. In sharp contrast, Almatrafi, Johri and Lee conducted a review of 47 articles written in 2019-2023, pointing to the diverse populations of learners involved, the various ways in which this construct was implemented, and how it was assessed.

The complexity of the concept of AI literacy is especially important in terms of higher education, since university students need not only to communicate with AI but also to evaluate information, generate knowledge, solve complicated tasks, and make decisions independently [9,10]. Therefore, AI literacy implies not only technical skills with AI systems but also the understanding of their functioning and limits, evaluating generated information, and deciding whether AI should be used in a particular task and, if so, how it can be applied. Thus, this broad concept underlies the choice of the scenario-based conception of AI literacy, which is considered in the current study. A recent synthesis of 29 studies adds weight to the integration of competencies in the concept of AI literacy and proposes a framework including the mentioned competencies for all kinds of learners and educational levels, making the idea of the uniform development path for AI-literacy competencies even stronger [11]. Nevertheless, there are some indications that the notion of AI literacy is also fragmented both conceptually and in practical terms. Whereas the reference framework defines the following relatively stable components as comprehension, application, evaluation, and ethical consideration, the modern framework includes such components as creation, social, integrational, legal, metacognitive, and contextual ones. It is a positive development, but it also raises the following issue: the increase in the number of components does not facilitate their practical application in the AI-based decision-making process.

B. AI Literacy as a Multidimensional Competence

Other forms of multidimensionality in AI literacy can be found in studies concerning measurement instruments of AI literacy. In this regard, the systematic literature review conducted by Lintner has revealed 22 papers containing validation results for 16 scales of AI literacy. While most of the reviewed scales revealed good structural validity and internal consistency, there were certain limitations with regard to their measurement properties. Notably, only a few scales were evaluated for content validity, reliability, construct validity, and responsiveness; no cross-cultural validity or measurement error was reported. More critically, only three of the reviewed scales were performance-based while thirteen were self-report scales.

The issue of the mismatch between self-report and performance-based measures is especially pertinent to the current toolkit proposal. If learners are asked about their understanding of AI, about their ability to properly trust AI and use it, then this assessment will measure the learner's

perception of their own capabilities rather than the demonstrated competence. The discrepancy between self-assessment and performance-based assessments of AI literacy has been recently established as well. In light of that, the current study defines AI literacy as a capability which must be ultimately measured by action. That is to say that the learner cannot be considered fully AI literate just because they can describe the concepts of hallucination, bias, and automation bias. The learner must demonstrate that they are able to detect these issues in a real-life task and react to them adequately.

Evidence quality is evaluated in terms of study design and context, not by the number of sources alone. A systematic review can identify recurring competencies, but it cannot demonstrate that a newly proposed Defense-specific instrument is valid. Similarly, results obtained from small interventions or Defense studies make useful comparison cases, but they are not direct applications of the tools at hand.

Table II contrasts concept reviews, measurement tools, small educational interventions, and Defense-related behavioral findings. It outlines their unique value without suggesting that results obtained from one population can be generalized to another.

Table II. Evidence on AI-Literacy Conceptualization and Measurement:

Study	Year	Evidence/design	Sample/corpus	Principal contribution	Relevance to toolkit
Almatrafi et al.	2024	Systematic review	47 articles (2019–2023)	AI-literacy constructs, implementation and assessment	Competency synthesis; not an intervention test
Lintner	2024	Systematic review	22 studies / 16 scales	Measurement quality and limitations	Supports performance-based assessment
AI-literacy measurement review	2024	Scale analysis	Multiple instruments	Four correlated dimensions identified across instruments	Supports multidimensional structure
Yan et al.	2025	Intervention	16 students	Knowledge, metacognition, ethics, and contextual AI use	Supports educational intervention design
Kahn et al.	2026	Survey experiment	236 cadets	Automation bias, algorithm aversion, and calibrated trust	Provides a defense-relevant behavioral dimension

There are three mutually complementary types of evidence in the comparison: the systematic review of Almatrafi et al., who analyzed 47 papers on AI literacy competence and described its core components among different populations; the analysis by Lintner of 16 scales used to measure 22 validation studies and their high reliance on self-reporting; the intervention literature demonstrating possible ways of teaching; and finally, the cadets' literature addressing the issue of reliable reliance among a Defense population. All four types together prove that it is reasonable to consider the definition of competence and human–AI judgment.

C. Generative AI Literacy in Higher Education

The emergence of generative AI technology has added new requirements that cannot be met by traditional definitions of AI literacy. GenAI can create fluent and contextually relevant texts, thus making one of the main challenges in learning about GenAI the following one: An output

created by a system may look reliable while the facts, reasoning, sources, or assumptions behind it are incorrect [12,13]. Thus, some researchers have started to make a distinction between GenAI literacy and the more general concept of AI literacy. In 2025, there was a systematic review of literature on GenAI literacy that identified five competencies in GenAI literacy: knowing and understanding GenAI; using and applying GenAI; evaluating and incorporating GenAI; GenAI ethics; and attitudes towards GenAI. It included the analysis of 51 empirical studies published in 2023 and 2024, proving how fast the research on GenAI literacy has developed. In addition, a systematic review done in 2025 and aimed at studying the relationship between GenAI and academic skills includes 158 studies published in 2021-2024, showing how actively scientists explore the impact of GenAI on cognitive, technical, and interpersonal skills in higher education.

From these data, it is evident that GenAI literacy cannot be considered just an addition to prompt-writing skills. Rather, learners need to develop the skills that will help them to evaluate generated outputs, recognize the situations in which the outputs can be unreliable, appropriately combine AI help with reasoning, and ensure academic and professional responsibility. This fact is particularly true for Defense Higher Education since the consequences of an apparently reasonable but actually incorrect output may be much worse than in other contexts. Also, one of the five areas of under-researched governance issues in generative AI ethics in education is the issue of human accountability for AI decisions [14].

D. Critical Evaluation and Verification of AI Outputs

Critical assessment of the information produced by AI technologies is one of the essential aspects of AI literacy. There is a possibility that the information produced by an AI algorithm might be linguistically convincing yet not accurate, unsupported, fictional, incomplete, or based on incorrect assumptions.

Therefore, the ability to verify should be one of the explicit parts of AI literacy. The information provided in the references shows that evaluation of AI literacy should take into account knowledge, skills, explanations, source verification, and decision responses instead of just confidence. In the context of Defense education, this distinction becomes crucial because students have to deal with information that is incomplete, uncertain, controversial, or consequential. Verification in this case means:

1. What is the claim?
2. What evidence exists to back up the claim?
3. Where did the evidence come from?
4. Is the evidence source credible and suitable?
5. What is still uncertain?

Empirical studies in the field of professional education, which is considered a high-risk setting, shed light on the problems caused by using unverified explanations from artificial intelligence: a randomized experiment conducted on inexperienced medical students showed that misleading explanations from artificial intelligence led to decreased diagnostic ability when compared with the no-explanation control group, and there was no difference in confidence between the two groups, hence a lack of confidence calibration [15].

E. Human Agency and Critical Judgement

AI literacy is also highly linked with human agency. Human agency deals with whether the learners maintain control of their reasoning, intentionality, authorship, and responsibility during the use of AI, while critical judgment focuses on learners' capacity to assess claims, evidence, uncertainty, bias, and consequences. While related constructs, both agency and

critical judgment are not the same and must not be confused. Reference evidence shows that interactions between learners and AI can be structured in such a way that would either promote or limit human agency [16]. In addition, reference evidence states that learning interventions may help to develop metacognitive skills, critical judgment, ethics, and context-based application of AI technologies in conjunction with technical skills. This difference is particularly significant in terms of Defense Higher Education, as AI-assisted decisions must not be equated with AI-motivated decisions. In the appropriate situation, a learner can use AI in order to produce different options and organize information, remaining responsible for their evaluation and final decision.

F. Calibrated Trust, Reliance, and Automation Bias

There is also a body of literature on the link between AI trust and reliance. Trust does not need to be maximized but must be calibrated to match the system's capability, quality of evidence, level of uncertainty, cost of making the wrong decision, and the possible consequences of failure. Literature on the issue clearly defines trust as an attitude and reliance as a task-specific action [17]. It is especially important for Defense education, where two mistakes can occur. Automation bias makes one trust the AI too much, while algorithm aversion makes one reject its help without a good reason [18].

Literature that includes military academy cadets is especially relevant as it investigates automation bias, algorithm aversion, and calibration of trust in Defense students [19]. However,

The above-mentioned reference paper properly highlights that there is reason to be careful with the results of the study due to its nature as a preprint with several shortcomings. Thus, the presented toolkit does not regard "trust in AI" as a result of education on its own. The tool is focused on the process of choosing the proper response to the use of the technology: Accept → Question → Verify → Modify → Override → Reject. This allows for a more behavioral approach to the concept of responsible AI literacy. Military evidence complementing the above-mentioned studies shows that trust in the use of AI for the purposes of decision-making at the strategic level is driven by the balance of numerous technical, operational, and oversight factors [20]. Another study conducted among cadets showed that the assumption about trusting attitude towards AI technologies as "digital natives" was not true [21].

G. Scenario-Based Learning and Performance Assessment

Self-report measure limitations give an excellent theoretical explanation for why scenario learning should be encouraged. Scenarios give education scholars an opportunity to examine how learners react to changes in context. It is possible that the same AI system will generate useful outputs in a certain non-threatening scenario and yet needs considerable validation in a different threatening scenario. Scenario-based evaluation can thus manipulate:

- AI-system reliability;
- uncertainty;
- evidence quality;
- task stakes;
- conflicting information;
- time pressure;
- institutional rules; and
- consequences of error.

Students may then be encouraged to justify their choices, to provide the evidential base they relied on, to account for any discrepancies, to identify the appropriate dependence response,

and to determine who is responsible for the eventual outcomes. This is consistent with the advice provided in the reference framework for performance assessment to ask students to take on responsibilities like accepting, challenging, verifying, revising, overriding, and rejecting the outputs of the AI system in justifying their decisions. Evidence from professional learning situations other than AI learning provides additional support for such an approach [22].

H. Defense Higher Education as a High-Stakes Transfer Context

The Defense context imposes one more theoretical consideration: the transfer needs to be context-specific. General higher education literature may prove helpful in providing evidence about AI literacy mechanisms; however, it cannot automatically ensure their validity in the case of Defense education. The reference literature makes it clear that transfer is contingent upon the population involved, the stakes of the task, authority relations, assessment settings, institutional governance, and consequence profiles. It allows making the distinction between AI literacy generally and AI literacy in the context of the Defense organization. The former could involve knowing about AI, applying AI technologies, assessing outcomes, and acting responsibly. The latter, besides all that, should take into consideration professional responsibility, high-stakes decision-making, institutional accountability, uncertainty, and the necessity to question AI applications when evidence is not convincing enough. The policy literature of Defense can provide the governance standards and institutional demands [23,24,25]; the empirical literature of higher education can provide information about the learners' capabilities and educational mechanisms [26]. These evidence categories should not, however, be treated as interchangeable. The reference paper emphasizes that policy expectations do not demonstrate learner attainment, while general higher-education evidence does not automatically demonstrate Defense learner capability. Defense-focused work on secure and explainable language models illustrates a relevant application context, although it does not itself validate an AI-literacy assessment toolkit [27].

Table III provides a structured summary relevant to this section. The table connects AI-literacy constructs to illustrative scenario challenges, observable learner actions, and proposed assessment indicators.

Table III. Illustrative Scenario-Based Operationalization of AI-Literacy Constructs Identified in the Literature:

AI-literacy construct	Scenario challenge	Observable learner behavior	Potential assessment indicator
AI understanding	AI produces a confident but uncertain response	Identifies system limitations	Accuracy of limitation identification
Critical evaluation	AI provides competing claims	Challenges unsupported claims	Quality of reasoning
Verification	AI provides factual assertions	Checks authoritative sources	Verification completeness
Reliability judgement	AI output varies in quality	Recognises uncertainty	Calibration accuracy
Reliance	Evidence is strong/weak	Accepts or questions appropriately	Appropriateness of reliance
Human agency	AI recommends a decision	Retains independent judgement	Justification of final decision
Ethics/accountability	AI output creates an ethical concern or contributes to final work	Identifies ethical implications; documents AI contribution	Ethical reasoning; accountability statement

The table Above highlights how the framework maps the constructs of AI literacy to observable actions and performance indicators of the learners. It assesses the learners' capabilities of comprehending the limitations of AI, critically analyzing the claims and counterclaims, validating

the information by using authoritative sources, dealing with uncertainty, and calibrating the use of recommendations given by AI technologies. In addition, it also highlights the role of human agency and ethics in making assessments by measuring the capabilities of maintaining independent judgment, justifying the decision, and giving credit to the contributions of AI to learners.

I. Synthesis of the Literature and Implications for Toolkit Design

In summary, the literature suggests that AI literacy has evolved into a concept that is becoming more focused on capability, involving the understanding of, application of, evaluation of, ethicality of, reflection upon, human agency in relation to, verification of, and contextualized judgement about AI systems. Nevertheless, several limitations of the current body of literature should be noted: AI-literacy measurement remains largely self-reported, objective performance testing is rather underdeveloped, and transfer of general higher education results to the Defense domain requires further validation. The above literature informs the rationale for the proposed Scenario-Based AI Literacy Toolkit.

In the conceptual framework employed in the current study, AI literacy is viewed separately from responsible AI judgement, with the former providing knowledge and the evaluative capabilities that shape the latter. In other words, the former is a competence required by the latter, which is, however, not necessarily generated by the former and thus has to be observed. In contrast to simply replicating the existing competency frameworks, the current toolkit attempts to translate the key dimensions of the concept of AI literacy into decision-making behaviors under certain contexts.

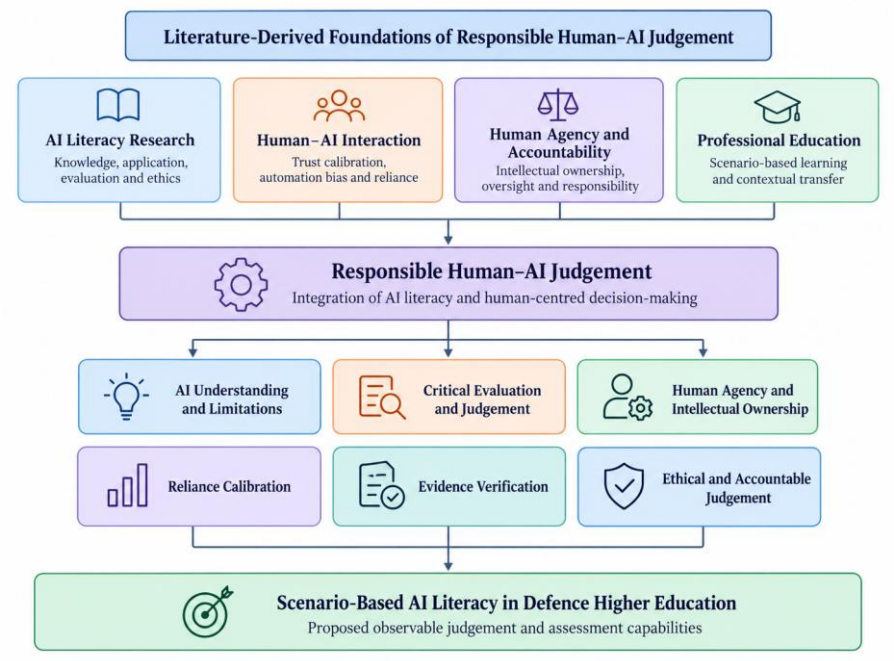


Figure 2. Literature-derived foundations and six provisional dimensions of responsible human–AI judgement.

As illustrated in Figure 2, there are four literatures that relate to the notion of responsible human AI judgement: AI literacy, Human-AI Interaction, Agency and Accountability, and Professional Education. The six provisional toolkit dimensions provide an organizational framework of these literatures in a way that can be used for educational purposes. They do not represent validated constructs or a developmental progression.

III. METHODOLOGY

This subsection presents the staged approach of the methodology that has been employed for the development of the Scenario-Based AI Literacy Toolkit for Defense Higher Education. The processes of searching and screening the evidence, the standards for synthesis of the evidence for AI literacy and human-AI judgement research, and its translation into the scenario-based teaching and assessment tools are described. Additionally, the difference between the design processes based on evidence, which has been performed within this study, and validation processes proposed for future studies is highlighted.

A. Research Design

For the purposes of developing a Scenario-Based AI Literacy Toolkit for Defense Higher Education, this study utilizes a two-part, staged research design. The first component consists of an initial evidence search and synthesis with characteristics akin to scoping research, wherein the ways that concepts such as AI literacy, generative AI literacy, critical evaluation, verification, human agency, balanced use, automation bias, accountability, and scenario-based assessment have been theorized and put into practice in higher education, professional education, high-stakes decision-making, and Defence/PME are mapped out. The second component will involve a prospective validation process by which the toolkit developed based on evidence is tested via expert review, scenario piloting, performance assessment, and psychometrics.

This has been done in such a way that evidence synthesis would inform the constructs, dimensions of the scenarios, and the assessment criteria, whereas validation would occur at a later point in time as part of the subsequent research process. It should be stated that the proposed toolkit is neither reliable nor validated in this particular research effort. It has been done in the same manner as in the reference study.

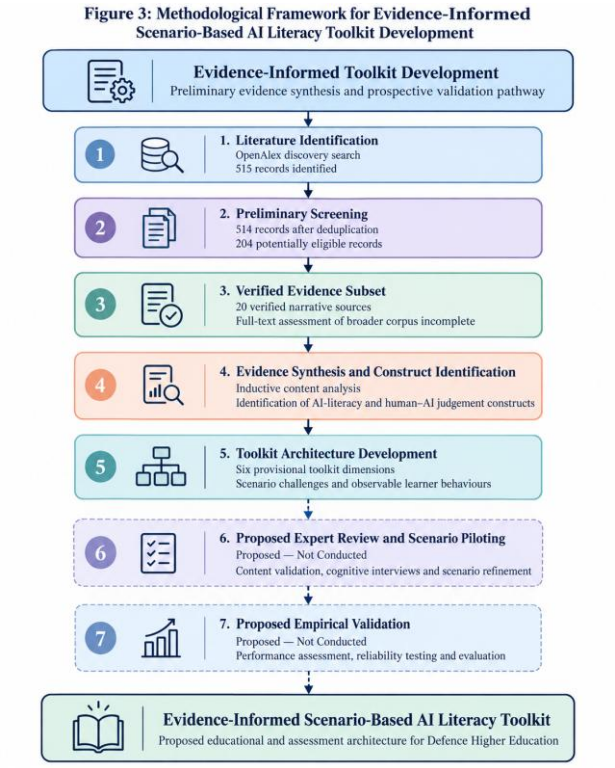


Figure 3. Evidence-informed toolkit development and proposed future validation stages.

As illustrated in Figure 3 above, the research was carried out following preliminary discovery, title and abstract screening, analysis of verified narrative sources, and development of toolkit

architecture. The expert review, piloting of the scenarios, reliability assessment, and empirical validation have been planned but not yet carried out.

B. Research Questions and Prospective Validation Hypotheses

The research questions (RQ1–RQ4) guide the present evidence-informed design study and are addressed directly by the evidence synthesis reported below. Hypotheses H1–H5, by contrast, are prospective and are reserved for the subsequent empirical validation study; they are not tested against data in the present manuscript. Question RQ1 deals with the concept and measurement of AI literacy and generative AI literacy. Question RQ2 deals with critical evaluation, human agency, verification, balanced reliance, automation bias, ethics, accountability, and scenario-based testing for learning and decision-making with AI. RQ3 addresses direct Defence/PME evidence and the bounded transfer of findings from higher education and other high-stakes professional contexts to Defense Higher Education. RQ4 addresses the derivation, design, and prospective validation of the Scenario-Based AI Literacy Toolkit, including the translation of evidence-informed constructs into scenarios, observable learner behaviors, and performance-based assessment indicators. The empirical validation protocol will test the following hypotheses after data collection; they are stated here as prospective hypotheses only and are not empirical findings of the present manuscript:

H1. Higher AI-literacy knowledge and awareness of AI capabilities and limitations will be positively associated with scenario-based judgement accuracy.

H2. Higher critical-evaluation and human-agency scores will be positively associated with verification quality and intellectual-ownership scores.

H3. Higher reliance-calibration scores will be associated with lower rates of both over-reliance and under-reliance when scenario reliability and task stakes vary.

H4. Participants receiving the structured Scenario-Based AI Literacy Toolkit intervention will show greater improvement in scenario-based judgement accuracy and verification quality than participants receiving a tool-orientation comparison activity, with allocation, sample size, and treatment fidelity to be determined by an a priori power analysis in the subsequent validation study.

H5. The correlation between AI literacy and accuracy of judgement in scenario-based tasks is expected to be partially mediated by quality of verification to some extent, under the assumption of sufficient sample size and proper modeling.

All these are prospective hypotheses only and not validated with any actual data, statistics, or validations in this paper.

Table IV shows a table related to this section. This table states the prospective hypotheses along with their requirements for evidence; none of which are tested statistically or proven.

Table IV. Hypothesis Audit Matrix (Prospective Not Yet Tested):

Hyp.	Theoretical rationale	Predictor	Outcome	Planned analysis	Principal limitation
H1	Understanding AI limitations should improve evidence-sensitive evaluation.	Objective AI-literacy score	Scenario judgement accuracy	Regression / mixed-effects model	Association ≠ causation

Hyp.	Theoretical rationale	Predictor	Outcome	Planned analysis	Principal limitation
H2	Agency and critical evaluation should support active verification.	Agency + critical evaluation	Verification quality + ownership	Multiple / mixed-effects regression	A construct distinction must be established.
H3	Appropriate reliance should change with reliability and stakes.	Calibration score	Over-/under-reliance	Scenario-level mixed-effects model	Requires sufficient observations
H4	Structured judgement training should outperform tool orientation alone	Toolkit vs comparison condition	Pre/post judgement + verification	Mixed-effects model	Requires valid allocation and treatment fidelity
H5	Verification may explain part of the literacy–judgement relationship	AI literacy → verification	Judgement accuracy	Exploratory mediation	Requires adequate sample and temporal ordering

C. Search Sources, Dates, and Access Boundary

The formal search for evidence will be carried out using Scopus, Web of Science, ERIC, IEEE Xplore, ACM Digital Library, and pertinent Defence/NATO/government repositories. These databases have been chosen based on the interdisciplinary nature of the topic in education, computing, human-AI interaction, psychology, professional learning, Defense studies, and governance. The time span covered by publication will be between 2019 and 2026, as this is a period of great advancements in the field of contemporary AI literacy and GenAI studies. Earlier methodological or theoretical papers can be included in the final literature list if they form the basis of constructs or review methodology used in the current study. At this point, the current paper does not claim that any particular database has been searched, as this has not yet taken place; Section IV identifies the single database searched to date (OpenAlex) from the others listed in this section that are intended for future formal searches.

D. Search Strategy and Search Strings

A structured search strategy was developed to identify literature relevant to AI literacy and the design of a Scenario-Based AI Literacy Toolkit for Defense Higher Education. The search targets Scopus, Web of Science, ERIC, IEEE Xplore, ACM Digital Library, and relevant Defense, government, and professional education repositories, using a combination of controlled keywords, Boolean operators, truncation, and phrase searching. The search concentrates mostly on articles produced from 2019 to 2026; however, earlier works, which form the basis of knowledge in the field, can be kept for their conceptual/methodological value. The search terms include AI literacy, generative AI, performance-based testing, human agency, verification, calibrated reliance, automation bias, and Defence/PME education.

Search terms were organised into four interconnected concept blocks: (1) AI literacy, including AI literacy, artificial intelligence literacy, AI competence, and AI competencies; (2) generative AI, including generative AI, GenAI, large language model*, and ChatGPT; (3) responsible AI-supported judgement, including critical evaluation, critical judgement, human agency, verification, trust calibration, reliance, appropriate reliance, automation bias, accountability,

and ethical judgement; and (4) educational and Defence context, including higher education, university, professional education, military education, Defence education, professional military education, and PME. These blocks are combined using Boolean operators to increase both sensitivity and specificity.

One of the main strings to be used for searching was developed as follows: ("AI literacy" OR "artificial intelligence literacy" OR "AI competence") AND ("generative AI" OR GenAI OR "large language model" OR ChatGPT) AND ("critical judgement" OR "critical evaluation" OR "human agency" OR verification OR "trust calibration" OR reliance OR "automation bias" OR accountability) AND ("higher education" OR "professional education" OR "military education" OR "Defense education" OR "professional military education" OR PME). Modifications in accordance with the specifics of databases searched and reference list searching and citation chaining will be performed where necessary.

The search will also involve synonyms such as defence/defense, military education, staff college, command and staff education, professional learning, human-autonomy teaming, machine advice, contestability, epistemic vigilance, and algorithmic reliance, as a search based on phraseology can ignore relevant articles that use different terminology.

E. Eligibility Criteria

Studies will be considered eligible when they directly or substantively contribute to understanding AI literacy, generative AI literacy, critical evaluation, human agency, verification, calibrated reliance, automation bias, ethical judgement, accountability, or scenario-based assessment within higher education, professional learning, high-stakes decision-making, or Defence/PME. The eligibility criteria will be applied consistently during title/abstract and full-text screening. Eligibility screening involves distinguishing eligible higher education and Defense-related research literature from technical AI literature that is not related to this context, opinion-based literature without support, duplicate literature, and literature that provides insufficient methodological information. Studies are eligible when they are: (1) published primarily from 2019 to 2026 but including earlier literature that provides a foundation for important theoretical or methodological concepts; (2) focused on higher education students, educators, professional learners, military/Defence learners or similar high-stake professional learners; (3) address AI literacy, GenAI literacy, AI-enabled judgements, human-AI interactions, critical thinking, verification, trust/reliability, human agency, ethics, accountability, or similar learning outcomes; (4) provide empirical, conceptual, methodological, review or authoritative policy/governance literature that is pertinent to developing the proposed toolkit; and (5) provide sufficient information about the research context, constructs, methodology, or main findings to allow for valid evidence extraction.

Articles will be excluded if they: (1) consider AI technology development or its effectiveness without bearing on human learning or judgement; (2) consider applications of AI that have no relation to literacy, use, learning or judgement in any form; (3) contain only opinions or commentaries without any evidential basis; (4) are redundant articles or lack necessary information for evidence classification; or (5) discuss applications where there is no reasonable way to draw lessons to higher education or Defence/PME without considerable unjustified extrapolation. Policy and governance documents will be used as sources to set institutional standards and ensure institutional oversight and assurance but not as empirical evidence of learner competencies; results from general higher education and other professional environments will be applied to the Defense Higher Education context only if the transition context is justified.

F. Evidence Selection and Screening

The evidence selection process will follow a systematic, multi-step screening process that is aimed at ensuring transparency, consistency, and reproducibility. Once the database searches have been performed, the articles that have been obtained from the databases will be exported to a reference management program, and duplicates will be flagged. The remaining articles will then be screened against the eligibility criteria using the titles and abstracts. Whenever possible, two independent reviewers will perform title/abstract screening and full-text screening, and any discrepancies will be discussed; in cases where there is no agreement, a third reviewer will make the final decision. The manuscript of this review will provide the total number of records identified, number of duplicates eliminated, number of records screened, number of records excluded based on titles and abstracts, full texts reviewed, full texts excluded with reasons, and number of articles used in the analysis. A flow diagram following the PRISMA format will be developed once the screening process has been completed; numbers will not be inserted until screening has been completed.

G. Evidence Categories and Rigor

Five types of evidence, which are not interchangeable, will be considered to differentiate the origin, context, and function of the literature that will form the basis of the toolkit. The first tier is made up of empirical evidence related to Defence/PME and represents the most contextually sound foundation from which to make claims about learners and the education environment in the Defense context. The second tier is comprised of high-stakes empirical research involving judgment, proper reliance, trust calibration, verification, and automation bias. Tier 3 comprises general higher-education empirical and conceptual evidence concerning AI literacy, generative AI literacy, learning, assessment, and related educational mechanisms. Tier 4 consists of human factors and professional comparator evidence relating to human/AI interaction, trust, dependency, verification, agency, and decision-making. Tier 5 consists of policy, governance, institutional, Defense, and professional guidance material, primarily employed for establishing expectations with regard to oversight, assurance, accountability, and institutional responsibility rather than learner performance evidence.

Methodological rigor is additionally reported using four descriptive bands: High, Moderate, Limited, and Conceptual/Policy, reflecting design, transparency, evidentiary strength, and applicability rather than fabricated numerical quality scores.

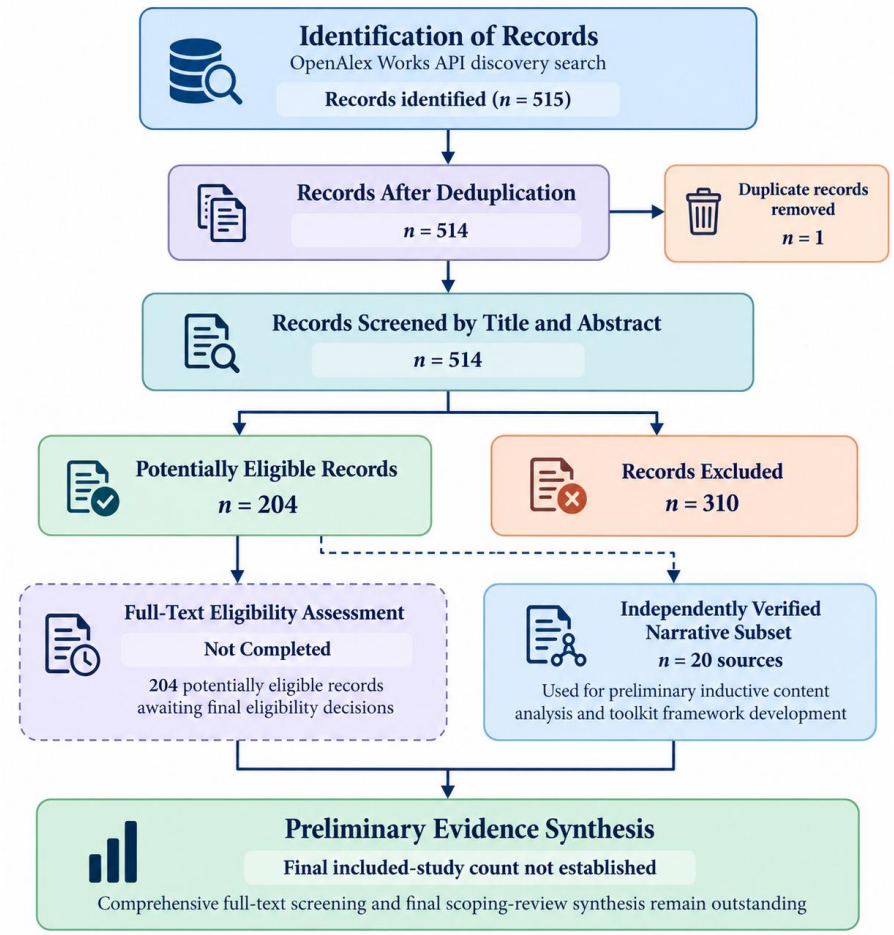


Figure 4. Preliminary OpenAlex evidence identification and screening; full-text eligibility remains incomplete.

As shown in Figure 4, the preliminary OpenAlex search returned 515 records; one duplicate was removed, leaving 514 titles and abstracts screened. Of those, 204 were potentially eligible and 310 excluded. Full-text eligibility decisions remain incomplete. The 20 independently verified narrative sources used for initial synthesis do not constitute the final included-study set.

H. Claim-to-Evidence Audit

In order to ensure the validity of claims and avoid exaggerated claims about the evidence basis, all the claims forming the toolkit will be subjected to an audit of the relationship between claim and evidence. The audit will assign an inferential category to each claim in terms of its status as an empirical fact, source-based claim, interpretation, bounded inference, hypothesis, policy implication, or proposed design aspect. Only those claims that are reported in the cited works in sufficiently comparable settings can be regarded as supported by evidence; support of one construct does not mean that the whole toolkit is validated.

Table V below is a structured summary related to this section. It differentiates source-supported claims, interpretations, and prospective designs to ensure that policy and evidence are not confused.

Table V. Claim-to-Evidence Audit Matrix:

Claim category	Evidence required	Example for present toolkit	Permitted interpretation
Empirical fact	Direct empirical finding	Learners demonstrate difficulty identifying AI limitations	Report as observed finding

Claim category	Evidence required	Example for present toolkit	Permitted interpretation
Source-based claim	Review/framework evidence	AI literacy includes evaluation beyond tool operation	Report as literature-supported
Interpretation	Compatible evidence from multiple sources	Verification appears important for responsible AI use	Present as interpretation
Bounded inference	Convergent evidence + explicit transfer assumptions	Verification mechanisms may be relevant to Defense HE	Label as inference
Hypothesis	Theoretical rationale requiring testing	Toolkit may improve judgement accuracy	Test empirically
Policy implication	Defence/institutional governance evidence	AI use should remain accountable to human oversight	Present as governance/design implication
Author proposal	Evidence-informed design rationale	Scenario-based verification task	Present as proposed toolkit feature

I. Framework and Toolkit Derivation

This Toolkit on AI Literacy using a Scenario-Based Approach was developed based on an evidence-to-design process rather than on the aggregation of dimensions of AI literacy as described in past frameworks. For the initial list of potential constructs identified during the evidence synthesis, the constructs were first mapped in terms of their recurrence, conceptual meaning, educational significance, observability, and appropriateness for Defense Higher Education. Next, each potential construct was evaluated against five decision criteria: (1) recurrence in independent pieces of evidence; (2) distinct conceptual meaning from other constructs; (3) observability via learners' actions or performance; (4) importance in terms of responsible judgement supported by AI; and (5) appropriateness for the Defense Higher Education setting.

See Table VI for further detail.

Table VI contains a detailed outline related to this section. The table reflects the thought process from research constructs to design features of the toolkit and distinguishes findings from source materials from the author's interpretation.

Table VI. Evidence-to-Toolkit Derivation Matrix:

Candidate construct	Recurrence	Uniqueness	Observable learner behavior	Resp. AI relevance	Defence transfer	Decision
AI Understanding and limitations	High	High	Identifies capability or limitation	High	High	KEEP
Critical evaluation/judgement	High	High	Challenges unsupported AI output	High	High	KEEP
Human agency/ownership	Moderate–High	High	Provides independent rationale	High	High	KEEP
Verification	Moderate–High	High	Checks sources and evidence	High	High	KEEP
Reliance calibration	Moderate–High	High	Adjusts reliance to reliability and stakes	High	High	KEEP

Candidate construct	Recurrence	Uniqueness	Observable learner behavior	Resp. AI relevance	Defence transfer	Decision
Ethical judgement	High	Moderate–High	Identifies ethical consequence	High	High	KEEP/ REFINE
Accountability/ oversight	High (governance /high-stakes evidence)	Moderate	Documents responsibility and escalation	High	Very High	MERGE/ REFINE
Well-being/ dependency awareness	Emerging	Moderate	Recognizes cognitive offloading and dependence	Context-dependent	Context-dependent	BOUNDARY CONDITION
Generic AI tool operation	High	Low	Uses AI functions	Moderate	Low–Moderate	MERGE/ EXCLUDE

J. Translation of Constructs into Toolkit Components

After the identification of constructs to be taught, each one of these constructs is turned into a scenario-based learning module through a four-step process, namely: Construct -> Scenario Context -> Observable Behavior -> Assessment Criteria. The operationalization of AI knowledge and awareness of limitations is done through scenarios where the AI generates information confidently that is not fully true or reliable, thereby forcing the learner to find out about the limitations of this AI system and what extra information it requires.

Verification is represented through source checking, discrepancy identification, provenance review, or escalation. Reliance calibration is assessed by varying AI reliability and task stakes and observing whether the learner appropriately accepts, questions, modifies, overrides, or rejects AI recommendations. Human agency and ownership are assessed through the learner's ability to provide an independent rationale and retain responsibility for the final judgement. Ethical and accountable judgement is operationalized through disclosure, role clarity, documentation, escalation, and responsibility for consequences, integrating the accountability/oversight and ethical-judgment candidates identified in Table VI into a single toolkit dimension, consistent with their MERGE/REFINE decision.

K. Final Toolkit Architecture

This final toolkit does not define the Defence/PME as an AI literacy competency. In its place, it is Defense Higher Education which serves as the contextual boundary that determines how the core competencies manifest and are measured in terms of the higher stakes, professionalism, institutional processes, assurance needs, uncertainty and auditability. The resulting framework is based on six core dimensions, in order as follows: (1) Understanding AI and AI's Limitations → (2) Critical Appraisal & Judgment → (3) Human Agency & Intellectual Responsibility → (4) Dependence Calibration → (5) Verification → (6) Ethical & Responsible Judgment, with Defense Higher Education running along the full pathway as the contextual boundary. This six-dimension framework is applied consistently through the rest of the manuscript.

The following Table VII represents the structure of this part of the manuscript. The table contains the six provisional core toolkit dimensions along with the associated scenario suggestions and indicators; the latter are not yet validated.

Table VII. Scenario-Based AI Literacy Toolkit: Six-Dimension Architecture and Assessment Framework:

Toolkit dimension	Scenario manipulation	Learner task	Observable indicator	Assessment evidence
1. AI understanding & limitations	AI produces a confident but incomplete output	Identify limitations	Correctly identifies capability boundary	Limitation-awareness score
2. Critical evaluation & judgement	AI output conflicts with credible evidence	Compare and evaluate claims	Challenges unsupported output	Judgement rubric
3. Human agency & intellectual ownership	AI recommends a decision	Make and justify final decision	Provides independent rationale	Ownership score
4. Reliance calibration	AI reliability and task stakes vary	Decide whether to accept/reject/modify AI advice	Adjusts reliance appropriately	Calibration score
5. Verification	AI produces a questionable claim/source.	Verify information	Checks evidence and records discrepancy	Verification score
6. Ethical and accountable judgement	AI output creates an ethical/bias concern or contributes to a consequential decision.	Identify consequences; document AI contribution and responsibility	Recognizes ethical implications; discloses, escalates, and accepts responsibility	Ethics rubric; accountability score

This architecture recognizes the context of Defence/PME, and not a distinct competency, as outlined in the evidence synthesis presented in Section II and the derivation process in Table VI.

L. Prospective Validation Protocol

Validation of the proposed tool will therefore involve a mixed-method approach that is designed to test for the content validity, usability, reliability, construct validity, and educational validity of the tool. Expert review of the proposed toolkit dimensions and scenarios will be done first, whereby experts in the fields of AI literacy, Defence/PME, educational assessment, and human-AI interactions will rate the relevance, clarity, contextual fit, and coverage of the dimensions and scenarios. Cognitive interview and pilot testing of the scenarios among learners will then follow, followed by a performance-based assessment stage, inter-rater reliability analysis, psychometric validation if possible (due to sample size and measurement assumptions), experimental or quasi-experimental design involving a comparison condition,

and longitudinal evaluation. The above validation process is in line with the methodology used in the literature, which recognizes the importance of expert review, cognitive piloting, scenario-based assessment, psychometric validation, and direct Defence/PME curriculum research prior to making any claims on validity and educational validity of a framework.

M. Methodological Boundary

In this study, the development of the evidence-based Scenario-Based AI Literacy Toolkit framework is performed as opposed to the development of an evidence-based valid psychometric measure or a demonstrated evidence-based education intervention. Literature evidence supports the selection of constructs of the toolkit, definition of conceptual boundaries of these constructs, and creation of scenarios that allow for observation of the relevant behaviors of learners; however, literature evidence alone is not sufficient for establishing the reliability, validity, efficacy, or educational influence of the toolkit. This boundary is particularly important because the evidence base identified to date is characterized by substantial reliance on self-report measures, limited objective or performance-based assessment, no Defense-specific AI-literacy scale identified within the verified evidence subset, and limited direct evidence from Defence/PME curricula.

IV. RESULTS

This section reports only what the preliminary evidence-selection process has actually established to date: the database search, deduplication, and preliminary eligibility screening, together with the narrative synthesis drawn from the independently verified subset of 20 sources. It does not report a completed systematic review, and it does not present conceptual synthesis as empirical outcome data. Findings are further distinguished throughout by evidentiary type: direct Defence/PME evidence, general higher-education evidence, evidence from other high-stakes professional contexts, and conceptual/author synthesis consistent with the evidence-category framework in Section 3.7.

A. Preliminary Evidence-Selection Status

The preliminary evidence search identified a broad but heterogeneous literature base relevant to the development of the proposed toolkit. Eleven Boolean search blocks were executed through the OpenAlex Works API on 22 August 2026, covering publications from January 2019 to August 2026. The search returned 515 unique records; DOI/title deduplication removed one duplicate, leaving 514 records for title/abstract screening. Of these, 204 records were identified as potentially eligible on the basis of their relevance to AI literacy, human–AI judgement, verification, reliance, agency, education, high-stakes decision-making, or Defence/PME contexts.

Full-text assessment of these 204 records has not been completed; these records therefore represent a candidate full-text set rather than a final included-study corpus, and no final included-study count or exclusion-reason log is reported. Within this boundary, 20 source records were independently verified for narrative synthesis, comprising systematic and scoping reviews, empirical higher-education studies, an intervention study, an objective scenario-based assessment study, human-factors and trust research, military decision-making evidence, PME practice evidence, and official Defence/NATO sources. As a matter of importance, the research done so far has been done only using the OpenAlex Works API, while other databases like Scopus, Web of Science, ERIC, IEEE Xplore, ACM Digital Library, as well as Defence/NATO/government archives specified in section 3.3 have been identified as sources to conduct a proper search at a later stage.

The evidence therefore provides a preliminary basis for toolkit construction but does not constitute a completed PRISMA-ScR review or establish a final included-study count. The search coverage must also be interpreted conservatively. The executed search reported in the present manuscript was conducted through the OpenAlex Works API only. Scopus, Web of Science, ERIC, IEEE Xplore, ACM Digital Library, and relevant Defense, NATO, government, and professional repositories identified in the methodology represent intended sources for a subsequent formal search and were not searched as part of the evidence-selection results reported here. Consequently, the current evidence base should not be interpreted as exhaustive coverage of the interdisciplinary literature.

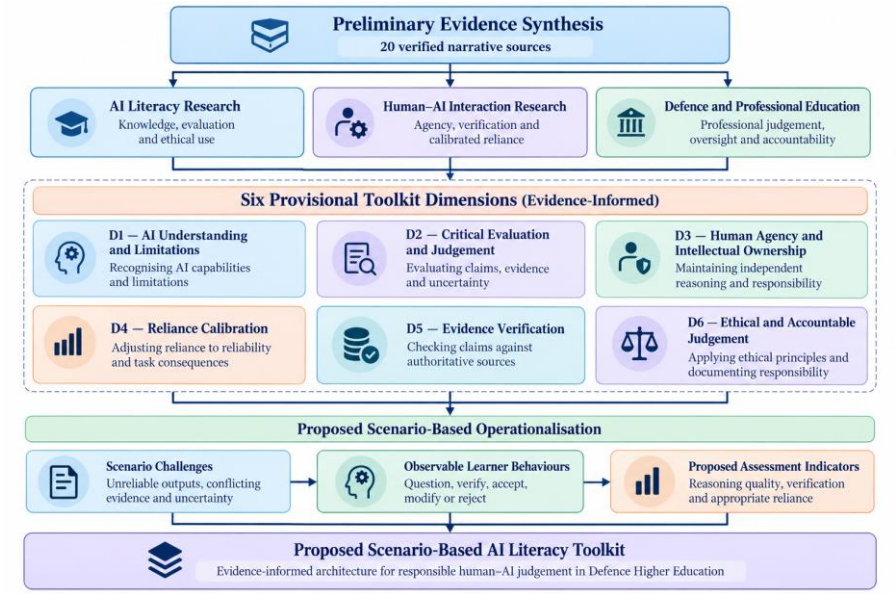


Figure 5. Evidence-informed development of six provisional toolkit dimensions and proposed assessment applications.

As shown in Figure 5, the preliminary 20-source narrative synthesis informed six proposed toolkit dimensions, which are linked to illustrative scenario challenges, observable behaviors, and potential assessment indicators.

These pathways represent the proposed operationalization of literature concepts, but do not constitute proof that there is validation of the toolkit, or that learner performance has been improved as a result.

Table VIII summarizes key information pertinent to this discussion. This table includes preliminary counts of records, as well as explicitly stating that the included study count remains undefined.

Table VIII. Preliminary Evidence-Selection Results:

Selection stage	Result	Interpretation
Boolean search blocks	11	Separate AI-literacy, judgement, trust, verification, and Defence/PME searches
Unique discovery records	515	OpenAlex discovery corpus
DOI/title duplicates removed	1	Transparent deduplication
Records after deduplication	514	Title/abstract screening set
Metadata/title-abstract exclusions	310	Did not meet AI + context + target-construct rule

Selection stage	Result	Interpretation
Potentially eligible records	204	Candidate full-text set

Selection stage	Result	Interpretation
Full-text reports assessed	Not completed	Access/verification boundary
Final included-study count	Not established	Must not be reported as a completed review result
Independently verified narrative subset	20	Used for preliminary narrative synthesis

It is vital that the difference between the 204 potential sources and the 20 validated narrative sources be made clear in the methodology. The 20 sources can be better termed a "subset of validated evidence" because they have been used to determine recurring constructs and evidence gaps in order to design the toolkit.

B. AI Literacy and Generative AI Literacy (General Higher-Education Evidence)

The evidence shows that the best way to conceptualize AI literacy is to look at it from a multidimensional perspective, which differs greatly from a one-dimensional understanding of tool proficiency. Throughout the examined literature, recurring dimensions of AI literacy include recognition of AI, understanding of AI, application and usage of AI, evaluation of AI products, creation with AI, and ethics of AI. Recent frameworks also highlight critical-thinking, social, integration, legal, meta-cognitive, and contextual dimensions. Given this, an ideal Defense Higher Education toolkit would have to go beyond teaching students how to operate certain tools and check if they understand what AI can and cannot do and how to properly evaluate the products of AI. The measurement evidence also highlights another limitation of the current approach to AI literacy: Self-report measures of AI literacy are prevalent, whereas objective performance measures vary in their validity. A recent study that used an objective measure of performance involved 531 undergraduate students and showed a very weak correlation between AI literacy and accurate classification of ethical scenarios. This evidence is general education and does not constitute any PME/Defence evidence.

Figure 6 summarises the relative magnitude of the evidence gaps identified within the verified subset (qualitative ratings, not measured effect sizes).

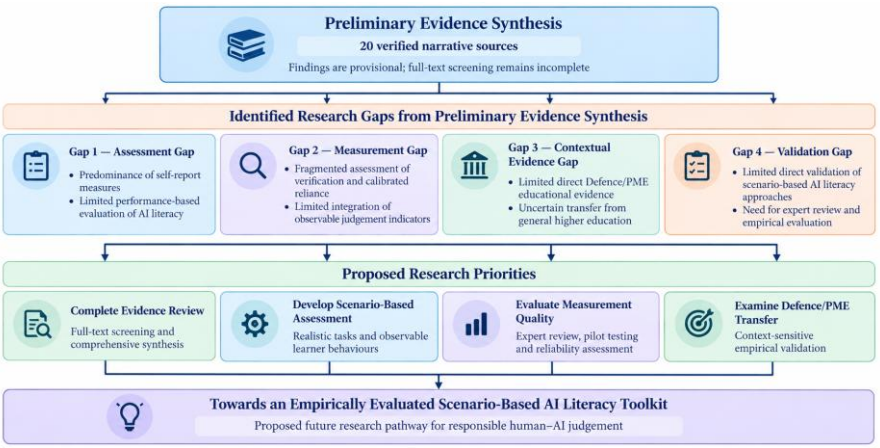


Figure 6. Preliminary research gaps and proposed priorities for scenario-based AI literacy research.

As shown in Figure 6, preliminary synthesis highlighted gaps in performance-based assessment, integrated measurement of verification and reliance, Defence/PME-specific educational evidence, and direct toolkit validation. These limitations motivate proposed future activities: completion of

the review, scenario development, measurement testing, and context-sensitive evaluation. The figure does not assign quantitative strength to any gap.

The illustrated gaps motivate direct Defence/PME testing and greater use of performance-based assessment.

C. Critical Judgement and Human Agency (General Higher-Education and Human-Factors Evidence)

The evidence indicates that human agency and critical judgement are related but distinguishable constructs. Human agency concerns the learner's retention of purpose, reasoning, authorship, and responsibility, whereas critical judgement concerns the evaluation of claims, evidence, uncertainty, bias, and consequences. Student–AI collaboration research indicates that the structure of AI interaction can either support or constrain agency, while intervention evidence suggests that metacognition, evaluation, ethics, and context-sensitive use can develop alongside technical tool competence. This result indicates that the successful use of AI should not be reduced to achieving the goal of generating an adequate response with AI for the purposes of the toolkit. Instead, the scenario activities have to require learners to analyze, question, and own the outputs supported by AI. Evidence about human-AI judgement helps to distinguish between attitude towards trust and real decisions regarding the use of AI.

D. Trust Calibration, Appropriate Reliance, and Verification (Human-Factors and Military Evidence)

The verified evidence distinguishes trust as an attitude or expectation from reliance as a task-level behavioral decision. Appropriate reliance should correspond to demonstrated AI capability, available evidence, uncertainty, task stakes, and potential consequences. Military and cadet evidence further indicates that both automation bias and algorithm aversion are relevant failure modes; a cadet study involving 236 cadets examined these patterns, although its preprint status and methodological limitations restrict the strength of the conclusions that can be drawn. Military decision-making evidence additionally emphasizes uncertainty, challenge mechanisms, contextual reasoning, and audit trails.

E. Defence/PME Evidence

The Defence/PME evidence provides strong contextual justification but limited direct educational validation. Military and Defense sources consistently emphasize human judgement, uncertainty, challenge mechanisms, audit trails, assurance, responsible AI use, and institutional accountability. Direct Defense evidence includes cadet research concerning automation bias and algorithm aversion, while PME practice literature identifies opportunities, limitations, ethics, and implementation considerations. These sources do not, however, establish that a particular Defense curriculum produces measurable gains in AI literacy, verification, reliance calibration, or accountability. The evidence therefore supports treating Defense Higher Education as the contextual boundary of the toolkit rather than as an additional AI-literacy competency; direct PME evidence remains comparatively limited.

F. Evidence Gaps

Four major evidence gaps emerged from the verified subset: (1) AI-literacy assessment remains heavily dependent on self-report, with limited objective and performance-based measurement; (2) no verified Defense-specific AI-literacy scale is established within the evidence subset; (3) education-facing measures of calibrated reliance and verification remain

fragmented; and (4) direct evidence demonstrating that a Defence/PME curriculum improves AI-literacy-related performance is currently insufficient.

Table IX provides a structured summary relevant to this section. The table groups the identified limitations of the preliminary evidence base and links them to the need for further assessment research.

Table IX. Evidence Gaps Relevant to Toolkit Development:

Evidence gap	Current status	Toolkit response
Self-report dominance	Substantial	Add performance-based assessment
Defense-specific measurement	Not established	Develop Defense-contextualized indicators
Verification measurement	No validated rubric identified	Create observable verification rubric
Reliance calibration	Education measures fragmented	Use scenario reliability/stakes manipulation
PME curriculum efficacy	Direct evidence limited	Treat toolkit efficacy as prospective
Transfer across Defense contexts.	Untested	Explicitly test contextual transfer

G. Evidence-Informed Toolkit Dimensions

The evidence synthesis supports a provisional set of core toolkit dimensions. AI understanding and limitation awareness and critical evaluation/judgement are retained because they recur across AI-literacy, ethical-scenario, intervention, military-decision, and Defense-policy evidence. Human agency and intellectual ownership are retained because the verified evidence directly links AI-supported collaboration with agency and responsibility.

Revised reliability calibration and validation are included as they are significant based on human factors, military, strategic, and governance evidence, whereas assessment in an educational context is inconsistent. Ethical judgment and accountability/oversight are combined under the one dimension, ethical and accountable judgment, due to Defense governance evidence suggesting a connection between authority, responsibility, ethics, and risk. Awareness of well-being/dependency is kept as a boundary condition rather than a main dimension of the toolkit, as no distinctive evidence about it is provided by the present literature base.

H. Evidence Synthesis and Toolkit Alignment

The findings of preliminary evidence have been synthesized in order to discover how the current literature can help shape the development of the proposed toolkit. Instead of considering all literature found as equivalent, it has been sorted according to construct, education context, population, function in methodology, and its applicability for Defense Higher Education. Direct Defence/PME literature has been separated from military and other high-stakes professional evidence, while higher education literature was mostly used as the means of establishing AI literacy and mechanisms of learning. Human factors and professional comparative studies were employed when appropriate, in cases where relevant evidence about trust, reliance, verification, automation bias, and human judgement was available.

These six constructs identified through the synthesis included AI literacy and limitation knowledge, critical appraisal and judgments, human agency and intellectual property, calibration, verification, and ethical judgments. They have not been chosen on the grounds of being present within individual pieces of literature; instead, these constructs were considered based on their recurring presence in evidence, their distinctiveness conceptually, the

possibility to observe them among learners, and relevance for responsible judgments supported by AI and justifiable transfer into Defense Higher Education institutions.

These constructs were then transformed into observable scenario-based behaviors to ensure that the toolkit assessed the learner's performance as opposed to only self-reported confidence or attitude levels: AI literacy and limitation awareness can be demonstrated through the identification of deficiencies in confident yet incomplete AI outcomes; critical judgment can be tested through discrepancies between AI suggestions and credible evidence; human agency can be measured through independent judgment and reasoning; reliance calibration can be evaluated using varying levels of reliability of the AI and importance of the tasks; verification skills can be demonstrated through source checking, discrepancy detection, and evidence documentation; and ethical and accountable judgment can be measured through responsibility assignment, disclosure, escalation, and oversight decisions. Evidence sources include AI-literacy literature, human-AI interaction studies, professional education, and defense/governance sources; the latter define the context requirements but do not provide evidence of learning outcomes.

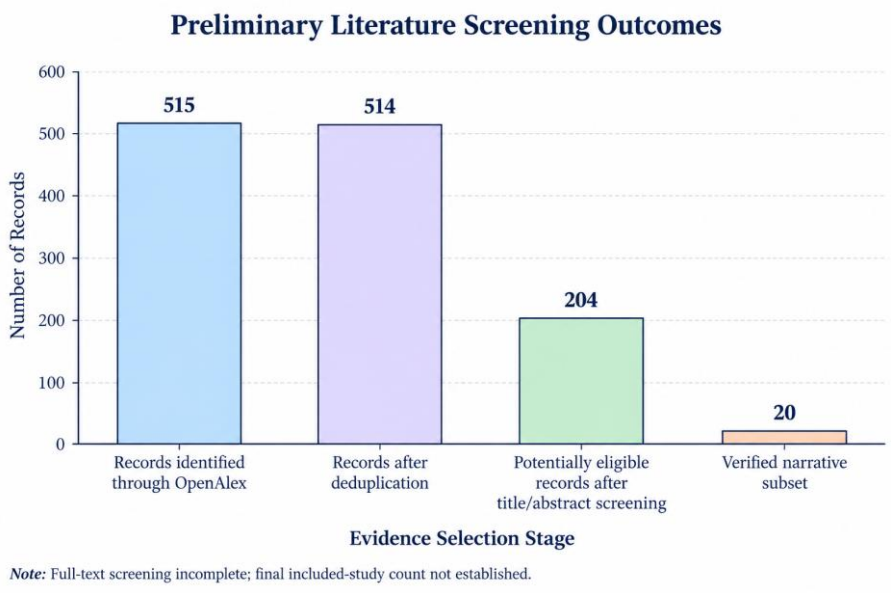


Figure 7. Preliminary literature screening outcomes, including the independently verified narrative subset.

Figure 7 shows that the selection process produced 515 records; 514 were duplicates removed, and 204 were classified as potential candidates based on titles and abstracts. These 20 narrative sources represent a distinct group of preliminary analysis sources instead of a final included study set. The bar chart depicts the number of records rather than a series of decisions.

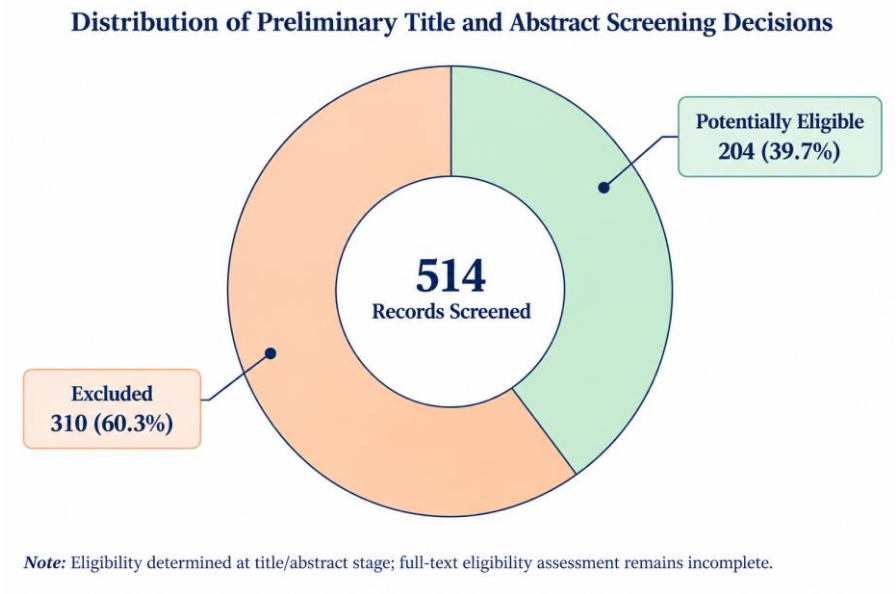


Figure 8. Distribution of preliminary title-and-abstract screening decisions among 514 records.

As illustrated in Figure 8, out of 514 records reviewed, 204 (39.7%) were identified as possibly eligible, whereas 310 (60.3%) were excluded based on title and abstract screening. The above statistics relate to the initial stages of screening, as the complete screening process has not been completed yet.

V. DISCUSSION

The architecture for the development of the evidence-informed Scenario-Based AI Literacy Toolkit for Defense Higher Education was established in the current research through incorporating evidence related to AI literacy, generative AI literacy, human-AI interaction, human agency, critical judgement, reliance, verification, and Defence/PME context. It can be seen from the findings that AI literacy cannot be considered just a technical skill or a measure of confidence in using AI applications. Rather, responsible usage of AI technology in Defense Higher Education requires a complex combination of competencies that enables individuals to be aware of AI limitations, be critical of outputs, retain intellectual autonomy, calibrate reliance on AI applications, verify information, and act in an accountable context. This is aligned with the evidence synthesis presented above but has limitations because the current research did not conduct a full-text assessment of all 204 potentially eligible studies. Thus, this paper cannot provide a systematically developed evidence corpus.

The primary contribution of this paper is the synthesis and contextualization of scenario-based AI literacy for Defense Higher Education – it is not the proposition that any of the constructs in question (critical evaluation, verification, human agency, appropriate trust, and accountability) have been new to the field of AI literacy. The constructs are all well documented in existing AI literacy, human factors, and military decision-making literature, but the contribution is in putting these constructs into a coherent judgment pathway and operationalizing them in terms of scenarios and assessment indicators.

A. What the Study Contributes

Implication one is that AI literacy and scenario-based judgement need to be seen as distinct but related outcomes. Evidence on AI literacy suggests conceptualization in terms of multi-dimensional concepts encompassing knowledge, understanding, using, evaluating, ethical considerations, and contextual application of AI systems. Yet the evidence on measurement

suggests that the dependence on self-report instruments persists while there is little evidence of objective assessment. This implies the distinction between the statement by the learner of being able to use AI on one hand and the demonstration of judgement on the part of the learner when an AI system offers incorrect, incomplete, or misleading information. Implication two pertains to the distinction that needs to be made regarding human agency, critical evaluation, reliance calibration, and verification. These mechanisms cannot be conflated into one measure of responsible AI use. Human agency relates to the capacity of the learner to retain reasoning and ownership and responsibility; critical evaluation refers to whether the AI outputs are critically evaluated in relation to the evidence, uncertainty, and potential consequences; reliance calibration refers to the extent to which the learner adjusts the reliance depending on capability, evidence, and task stakes; and verification refers to checking the information.

Military and human-factors evidence further indicates that both automation bias and algorithm aversion can create inappropriate reliance. Consequently, the toolkit does not reward simply accepting or rejecting AI; instead, scenario conditions require learners to determine when to accept, question, verify, modify, override, or reject an AI output and to justify the decision.

B. Why Scenario-Based Assessment Matters, and Relationship to Prior Frameworks

Existing AI-literacy frameworks (e.g.,) establish which competencies matter but do not, by themselves, specify how a learner's possession of those competencies should be observed under realistic conditions of uncertainty and consequence. The present toolkit does not compete with these frameworks; it operationalizes them. Scenario-based assessment matters because it is the only design examined in the

reviewed evidence that can distinguish a learner who merely reports confidence or knowledge from one who demonstrably adjusts behavior, accepting, questioning, verifying, modifying, overriding, or rejecting an AI output as reliability, evidence quality, and task stakes change. This is what conventional self-report AI-literacy instruments cannot capture, and it is why Defense Higher Education, where the cost of miscalibrated reliance is potentially high, is a particularly consequential setting in which to require it.

C. Transferability of Evidence from General Higher Education

A third implication concerns the Defense transfer boundary. General higher-education evidence can provide useful mechanisms and instructional hypotheses, while military and high-stakes professional evidence can clarify issues of uncertainty, reliance, automation bias, challenge mechanisms, and consequential judgement. Defense and governance evidence can establish expectations concerning oversight, assurance, accountability, and professional responsibility.

However, these streams of evidence cannot be considered equivalent. Transfer to Defense Higher Education is justified only if there is sufficient equivalence between the two on issues of population, task structure, authority relations, assessment circumstances, stakes, and consequences. Defense Higher Education must be considered, then, more a context boundary than another AI-literate skill.

D. Interpretation of the Toolkit Architecture

AI Understanding and Limitations, Critical Evaluation and Judgment, Human Agency and Ownership, Reliance Calibration, Verification, and Ethical and Accountable Judgement – This six-dimensional architecture of AI understanding and its limitations, critical evaluation and judgement, human agency and ownership, reliance calibration, verification, and ethical and accountable judgement is an educational design framework, not a causal model validated by

empirical research. The dimensions discussed should not at this point be treated as latent factors, as a fixed progression, or even as universal determinants of judgement quality. Instead, these are educational levers that can be tested using content validation, scenario-based performance measures, psychometrics, and replicability across contexts.

E. Contradictory Evidence and Measurement Implications

The findings also show inconsistencies that actually support rather than refute the logic of the toolkit. In self-report measures, individuals may show great levels of confidence, while in objective or scenario measures, they may exhibit a lack of clarity, ceiling effects, and difficulty in distinguishing between AI content and original content. Likewise, in trust literature, the dangers of over-reliance on AI and over-rejection of AI have been observed. Policy papers may outline responsible AI practices, while educational papers may rarely show that the principles mentioned in the policy papers result in learners' behavior. These inconsistencies show that it is not advisable to combine attitudes, knowledge, judgment performance, verification behavior, and accountability enactment into one composite measure. The future evaluation must consider the types of sources from which the findings have been obtained and their inferential distance.

F. Practical Implications for Defense Higher Education

What would appear to be the most justified use of such an AI tool for learning purposes at this point is conducting an unclassified scenario exercise with the learners being provided with an AI-supported analysis and then being tasked with recognizing the limitations of the system, analyzing the stakes of the task, comparing the AI's statements with relevant literature, etc.

Scenarios should include useful, misleading, incomplete, and ambiguous AI outputs so that learners cannot succeed simply by accepting or rejecting AI consistently. Assessment should focus on reasoning quality, provenance checking, verification, disclosure, escalation, and responsibility statements, rather than response speed, polished language, or agreement with the AI system. At the institutional level, the toolkit also implies alignment between curriculum, assessment, AI-use policy, and assurance arrangements. These are proposed educational implications, not evidence that existing Defence/PME systems already achieve these outcomes.

VI. LIMITATIONS

Several limitations should be considered when interpreting the findings of the present study.

First, the evidence synthesis is preliminary rather than a completed systematic or scoping review. Although the discovery search identified 515 unique records and 204 potentially eligible records, full-text eligibility assessment was not completed. The 20 independently verified sources therefore represent a narrative evidence subset rather than a final included-study corpus. Consequently, the present findings should not be interpreted as exhaustive estimates of the Defense Higher Education evidence base. Second, the evidence base is heterogeneous across populations, contexts, and methodological designs.

The existing literature involves general higher education literature, AI literacy frameworks, human factors research, military/professional literature, PME practice literature, and Defence/governance literature. Such sources cannot be considered to be interchangeable in terms of their evidentiary value. General higher education research can be used to guide educational mechanisms, while military/professional research can be used to guide reliance and automation bias mechanisms. However, neither necessarily indicates that learners in the context of Defense have achieved the desired AI literacy competencies.

Lastly, there is no doubt that the problem of measurement is significant. The research on the topic of AI literacy continues to rely on self-report measures to a large degree, and there is practically no performance measure at all. In addition, there is no mention of any Defense-specific AI-literacy scale among the validated ones, while education-focused validation and calibration measures are not unified.

Fourthly, the toolkit architectural structure does not have empirical validation. In fact, the six elements cannot be seen at this stage as independent latent variables, a linear developmental pathway, or a universal predictor of decision quality. Instead, the toolkit is made up of an evidence-based structure of mechanisms that need to be further validated, cognition tested, scenario tested, performance measured, reliability measured, and psychometrically validated.

Lastly, it must be emphasized that the current study is incapable of demonstrating any causation regarding the educational outcomes associated with the use of the proposed toolkit. There is no evidence at all that suggests the implementation of the toolkit can result in increased accuracy of judgments, enhanced verification, better calibration of reliability, and higher levels of accountability behavior. Therefore, the current study should be considered an evidence-informed toolkit framework.

VII. FUTURE WORK

Future research needs to shift from the current approach of evidence-informed architecture of the toolkit towards systematic empirical validation. The most important step here is completing the evidence selection process through the full-text assessment of the 204 potentially eligible records, justifying exclusion, and generating the final evidence set and PRISMA-like flow record. The finalized search protocol must be officially registered prior to any analysis of outcome data. The other step is to validate the toolkit constructs and scenarios with experts and learners. First of all, the six proposed constructs need to be reviewed by defense higher education experts and those who know AI literacy. Then cognitive interviews will assess interpretation of scenario instructions, AI responses, evidence needed, and choice of answers by learners.

Third, the validity and reliability of the toolkit need to be determined in future research. It would be important to develop scenario-based tasks that are related to each of the six dimensions. The assessment of performance should involve rubrics as well as, when needed, independent raters. Inter-rater reliability, internal consistency when needed, construct validity, criterion-related validity, and sensitivity to scenario and task reliability need to be determined. Fourth, a validation experiment should be conducted in which the Scenario-Based AI Literacy Toolkit is compared with some appropriate control condition, e.g., conventional orientation or instruction with respect to AI tools. Measures should include scenario-based judgement accuracy, verification quality, proper use, limitations recognition, human agency/ownership, ethical judgement, and accountability. If conditions are appropriate, mediation and moderation analysis can be performed to explore whether verification quality mediates the relationship between AI literacy and judgement accuracy and whether the effect varies depending on AI reliability and/or task stakes. Such analysis needs to be preregistered.

In conclusion, there should be future research into transfer in a longitudinal sense as well as in terms of transfer between contexts. Testing at different times might provide evidence of whether skills acquired through the toolkit are sustained over time as well as whether any improvements achieved during training will transfer into practical Defense Higher Education tasks.

VIII. CONCLUSION

In this research, a framework architecture for a Scenario-Based AI Literacy Toolkit in Defense Higher Education was proposed, considering the evidence regarding AI literacy, critical appraisal, human agency, calibration of reliance, verification, and ethical and accountable judgement. It is implied from the synthesis that responsible AI literacy in Defense Higher Education can not be effectively captured through technical AI knowledge and confidence in AI. Rather, responsible behavior should be demonstrated by learners having knowledge and understanding about AI's strengths and weaknesses, critical appraisal of AI-generated information, ownership of decisions, calibration of reliance according to reliability and level of stakes involved, verification of critical assertions, and working in an ethical and accountable context.

However, the greatest benefit of the research lies in the integration of these models into the architecture of scenario-based and observable performance for Defense Higher Education and not their novelty. These models distinguish between foundational knowledge in AI and the decision-making processes that such knowledge should help us achieve, regard dependence on others as conditional and not just something we increase or decrease, and make verification and accountability observable through the learner's behavior.

However, on the other hand, the results draw definite limits to claims that can currently be made. The preliminary evidence base provides support for the design rationale behind the toolkit but does not provide evidence of its reliability, construct validity, Defense-specific validity, and educational effectiveness. Objective assessment scarcity, lack of a Defense-specific AI-literacy scale among the verified evidence items, fragmented verification and reliance assessment, and the lack of direct evidence on the PME curriculum are important limitations. Any future research should proceed via expert validation, cognitive interviews, scenario piloting, performance-based assessment, inter-rater reliability tests, psychometric validation, comparative experiments, and longitudinal studies. The key empirical challenge is not to prove that Defense learners embrace responsible AI principles but to test whether they use appropriate judgement under conditions of varying AI reliability, uncertainty, and consequences of AI-assisted decision-making. In this respect, the toolkit offers a testable framework for developing responsible AI judgement based on AI-literacy principles in Defense Higher Education.

REFERENCES

[1] D. T. K. Ng, J. K. L. Leung, S. K. W. Chu, and M. S. Qiao. Conceptualizing AI literacy: An exploratory review. Computers and Education: Artificial Intelligence, 2:100041, 2021.

[2] O. Almatrafi, A. Johri, and H. Lee. A systematic review of AI literacy conceptualization, constructs, and implementation and assessment efforts (2019–2023). Computers & Education Open, 6:100173, 2024.

[3] R. Parasuraman and V. Riley. Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2):230–253, 1997.

[4] J. D. Lee and K. A. See. Trust in automation: Designing for appropriate reliance. Human Factors, 46(1):50–80, 2004.

[5] L. J. Skitka, K. L. Mosier, and M. Burdick. Does automation bias decision-making? International Journal of Human-Computer Studies, 51(5):991–1006, 1999.

[6] T. Lintner. A systematic review of AI literacy scales. npj Science of Learning, 9:50, 2024.

[7] A. Ali, Z. Farid, A. H. Al-kassem, Z. A. Khan, M. Qamer, K. F. Khan Ghouri, M. Al Sakhnani, and A. M. Momani, "Development and use of artificial intelligence in the defense sector," in 2023

International Conference on Business Analytics for Technology and Security (ICBATS), 2023, doi: 10.1109/ICBATS57792.2023.10111113.

[8] D. Long and B. Magerko. What is AI literacy? Competencies and design considerations. In Proc. 2020 CHI Conf. Human Factors Comput. Syst., pages 1–16, 2020.

[9] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur. Systematic review of research on artificial intelligence applications in higher education—where are the educators? International Journal of Educational Technology in Higher Education, 16:39, 2019.

[10] W. Holmes, M. Bialik, and C. Fadel. Artificial Intelligence in Education: Promises and Implications for Teaching and Learning. Center for Curriculum Redesign, Boston, MA, 2019.

[11] H. Chee, S. Ahn, and J. Lee. A competency framework for AI literacy: Variations by different learner groups and an implied learning pathway. British Journal of Educational Technology, 56(5):2146–2182, 2025.

[12] E. Kasneci et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103:102274, 2023.

[13] J. G. Meyer, R. J. Urbanowicz, P. C. N. Martin, et al. ChatGPT and large language models in academia: Opportunities and challenges. BioData Mining, 16:20, 2023.

[14] M. I. I. Alras, A. Q. Emran, and A. M. Mohamed. Ethics and governance of generative AI in education: A systematic review on responsible adoption. Discover Education, 5:37, 2026.

[15] D. Teng, L. Tan, Q. Cao, et al. Impact of AI misinformation on diagnostic accuracy and confidence calibration in novice medical students. npj Digital Medicine, 9:356, 2026.

[16] M. H. Jarrahi. Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. Business Horizons, 61(4):577–586, 2018.

[17] R. Parasuraman, T. B. Sheridan, and C. D. Wickens. A model for types and levels of human interaction with automation. IEEE Trans. Syst., Man, Cybern. A, Syst. Humans, 30(3):286–297, 2000.

[18] K. L. Mosier, L. J. Skitka, S. Heers, and M. Burdick. Automation bias: Decision making and performance in high-tech cockpits. International Journal of Aviation Psychology, 8(1):47–63, 1998.

[19] L. Kahn, M. C. Horowitz, and L. Resnick Samotin. What is human in judgment? Comparing automation bias and algorithm aversion between the United States Military Academy and the general public. arXiv:2604.04333, 2026.

[20] P. Lushenko. AI, trust, and the war room: Evidence from a conjoint experiment in the US military. Cambridge Forum on AI: Law and Governance, 1:e52, 2025.

[21] P. Lushenko and R. Sparrow. Artificial intelligence and U.S. military cadets' attitudes about future war. Armed Forces & Society, 2024.

[22] S. Holdsworth, O. Sandri, J. Hayes, and S. Maslen. Scenario-based capability assessment in professional learning contexts. Journal of Adult and Continuing Education, 31(3), 2025.

[23] Ministry of Defense, UK. Defense artificial intelligence strategy. Technical report, UK Government, London, U.K., 2022.

[24] Ministry of Defense, UK. Ambitious, safe, responsible: Our approach to the delivery of AI-enabled capability in defense. Technical report, UK Government, London, U.K., 2022.

[25] U.S. Department of Defense. Responsible artificial intelligence strategy and implementation pathway. Technical report, Dept. of Defense, Washington, DC, 2022.

[26] C. A. Pfaff, C. J. Lowrance, B. M. Washburn, and B. A. Carey. Trusting AI: Integrating artificial intelligence into the army's professional expert knowledge. Technical report, US Army War College, Strategic Studies Institute, Carlisle, PA, 2023.

[27] A. Ali, H. T. Sheik, A. Raza, T. Hanif, S. G. Virk, and H. Riaz, "MilGPT: Secure and explainable large language model framework for military applications," in Proc. 11th International Conference on Next Generation Computing (ICNGC), 2025, pp. 65–68. Available: https://www.earticle.net/Article/A478462.

Declarations

Funding. No external funding was received.

Conflicts of interest. The authors declare no conflicts of interest.

Ethics and consent. Not applicable statement.

Data availability. Repository / access conditions, or not applicable.

Code availability. Repository / version, or not applicable.

Author contributions. CRediT roles and initials.

Acknowledgments. Optional. Delete if none.

Use of generative AI. Disclose AI tool use, or state that none was used.

ijsigma.org

SIGMA International Journal of Security, Intelligence, Governance & Modern AI

VOLUME 01 · ISSUE 01 · 2026 · ARTICLE 0002

RESEARCH ARTICLE OPEN ACCESS · ISSN: [ISSN to be assigned] · DOI: [DOI to be assigned]

A Leakage-Safe Evaluation Framework for Gated Autoencoder–Transformer Multi-Class Network Intrusion Detection

Azhar Javed1, Naik Nawaz2,* and Asher Mehfooz3

1 Information Security Department, National University of Sciences and Technology (NUST), Islamabad, Pakistan

2 MSCS, National University of Sciences and Technology (NUST), Pakistan

3 Department of Data Science, University of Kotli (UOK), Azad Jammu and Kashmir, Pakistan

Corresponding author: Naik Nawaz (naik.021nawaz@gmail.com) · ORCID: [0000-0000-0000-0000]

Author emails: Azhar Javed (ajaved.phdismcs@student.nust.edu.pk) Asher Mehfooz (asherzaki960@gmail.com)

ABSTRACT

Expressive representation learning together with leakage-safe evaluation is required to achieve reliable multi-class network intrusion detection. In this paper, we introduce a leakage-safe evaluation protocol for a gated autoencoder–Transformer architecture applied to seven traffic groups in the CICIDS2017 dataset. We use the partitioning method as an experimental condition: S1 reuses the stratified reference partitioning, S2 places exact duplicates into one partition and audits near duplicates, and S3 uses whole capture sessions or time partitions when a seven-class scenario is feasible. No preprocessing (filtering, imputation, scaling, class balancing, autoencoder training) is allowed on any test data. The provided architectural baseline achieves 99.61 ± 0.05% macro-F1 on a balanced, stratified subset of the data, but we could not verify its bibliographic history or duplicate/group-disjoint proof; we retain those figures as reference results, not as the outcome of the current framework. Contribution: An audit-ready protocol for separating the effects of architecture from data leakage via reference baselines, ablation studies, per-class metrics, and a reproducibility report.

Keywords: Network intrusion detection; multi-class classification; CICIDS2017; autoencoder; Transformer; gated feature fusion; data leakage; duplicate-aware evaluation.

Highlights

- Expressive representation learning together with leakage-safe evaluation is…
- In this paper, we introduce a leakage-safe evaluation protocol for a gated…
- We use the partitioning method as an experimental condition: S1 reuses the…

IMPACT STATEMENT

Expressive representation learning together with leakage-safe evaluation is required to achieve reliable multi-class network intrusion detection. In this paper, we introduce a leakage-safe evaluation protocol for a gated autoencoder–Transformer architecture applied to seven traffic groups in the CICIDS2017 dataset. We use the partitioning method as an experimental condition: S1 reuses the stratified reference partitioning, S2 places exact duplicates into one partition and audits near duplicates, and S3 uses whole capture sessions or time partitions when a seven-class scenario is feasible. No preprocessing (filtering, imputation, scaling, class balancing, autoencoder training) is allowed on any test data. The provided architectural baseline achieves 99.61 ± 0.05% macro-F1 on a balanced, stratified subset of the data, but we could not verify its bibliographic history or duplicate/group-disjoint proof; we retain those figures as reference results, not as the outcome of the current framework. Contribution: An audit-ready protocol for separating the effects of architecture from data…

I. INTRODUCTION

Multi-class Intrusion Detection Systems (IDS) analyze flows and distinguish normal traffic from attack traffic. Multi-class IDS classify the attack type, so different measures are required to counter brute-force attacks, denial-of-service attacks, port scans, and web attacks. CICIDS2017 contains class-labeled data and can be used for research. However, strong results on this dataset make sense only with complete documentation of the data source, its classification, and the evaluation process [1]. Randomly splitting the records does not mean the IDS performs well on independent network traffic.

A. Background

New approaches to intrusion detection rely on representation learning along with attention-based classification. The autoencoder encodes numerical flow descriptors into latent descriptors. At the same time, a transformer learns interactions among the feature descriptors for classification. Although raw and latent descriptors may contain different information (the first describes traffic characteristics, while the second captures nonlinear relationships), a learned gate can adjust their relative importance for each flow. This approach seems promising, but complexity alone does not guarantee the architecture's superiority. Strong baseline architectures exist for tree-structured tabular data [2]. Hence, a reasonable comparison would include such an experiment.

B. Problem Statement

The intrusion-detection system's performance may seem too good if the training and test datasets are not truly independent. Any normalization, feature selection, or autoencoder fitting before splitting will influence the model's training because it will learn from the final distribution. Although preprocessing is conducted correctly, exact or almost exact copies of flows from the same capture may be located in different parts of the split sample. This makes model performance evaluation more difficult. Data leakage research identifies contamination of model development and testing as a common problem in obtaining reliable predictions [3]. This problem is especially relevant in multi-class detection, where good performance can hide mistakes on less frequent attacks.

C. Research Gap

The provided architectural citation employs the traffic features in a standard format, frozen autoencoder embeddings, gated raw-latent fusion, a Transformer classifier, and focal loss training. On a balanced subset of 13,992 records across seven classes from the CICIDS2017 dataset, it achieves 99.61 ± 0.05% macro-F1. However, the paper provides no verifiable bibliographic citation for this material, and it lacks duplicate counts and session- and time-disjoint evaluations. Research on experimental bias shows how data collection and partitioning can significantly affect interpretations of security model performance [4]. Thus, this paper treats the provided score as a benchmark and asks whether the gated model retains its relative superiority when duplicate counts and session- and time-disjoint evaluations are considered.

Table 1: Existing evidence and the unresolved leakage-safe evaluation questions.

Evaluation issue	Evidence available	Unresolved question	Required analysis
Seven-class performance	Macro-F1 of 99.61 ± 0.05% Previously reported on the balanced stratified subset	Does the score persist with more independent test traffic?	Compare matched split protocols.
Preprocessing	Training-only fitting is described	Can the fitted objects and their input rows be independently audited?	Retain fit indices and preprocessing objects.
Duplicate flows	Cross-split counts are not supplied	Do identical or nearly identical flows cross partitions?	Exact-hash and near-duplicate audit
Capture relationships	Stratified random splitting is reported	Do related sessions or capture periods span splits?	Feasible group/time-disjoint holdouts
Minority attacks	Seven-class F1 scores are reported	How do minority-class errors change under stricter splits?	Per-class recall, PR analysis, confusion matrices
Model components	Reference-protocol ablations are reported	Do gating and focal-loss gains persist after the audit?	Matched ablations on audited partitions

Table 1 illustrates the separation between the evidence found regarding references and the unanswered questions. We have not yet done duplicate detection or time/partition-disjoint analyses for this paper. Figure 1 shows that the process starts with provenance and duplicate auditing, data partitioning, and fitting the transformation to the training data alone.

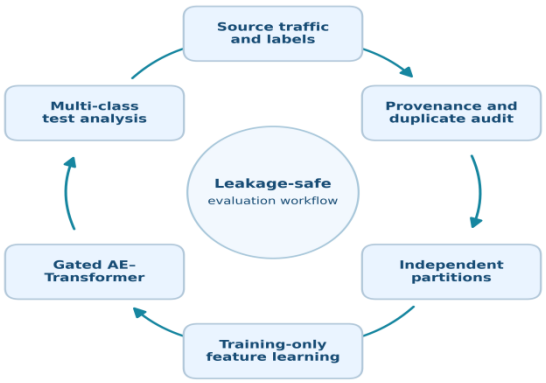


Figure 1. Evaluation methodology for a gated autoencoder-Transformer that is leakage-safe.

The six steps of the framework shown in Figure 1 are: source traffic and labeling, provenance and duplicate auditing, generation of disjoint partitions, learning transformations on training data alone, gated autoencoder-Transformer classification, and multi-class testing. This ordering matters because provenance and duplicates are checked before generating partitions, transformations are fit only on training data, validation is for model selection, and the last partition is not opened for evaluation until necessary.

D. Proposed Framework and Contributions

This article describes an auditable framework for evaluating gated autoencoder-Transformer on seven classes of CICIDS2017 in six steps: label and provenance checking, exact and near-duplicate checking, and comparing feasible session- or time-disjoint partitions with the stratified reference partition. In each partition, we train transformations only on training data, use validation data for model selection, and reserve test data for evaluation. This paper's contribution is methodological: the proposed design enables a fair comparison between baseline

classical and neural networks while examining the impact of latent representation learning, sample-dependent gating, Transformers, and focal loss across increasingly independent partitions.

E. Research Questions

This design framework centers on three testable research questions. RQ1 examines changes in total- and class-level macro-F1 and recall when moving from the less strict stratification S1 to the stricter S2 and feasible S3 partitioning. RQ2 asks whether the advantage of raw-latent gating and focal loss remains relevant as evaluation progresses from less strict to stricter partitions. RQ3 explores changes in minority-class behavior, false alerts, and efficiency as traffic independence increases.

II. LITERATURE REVIEW

Multi-class network intrusion detection has evolved from building classifiers based on single-flow features to latent representation learning, attention-based feature interactions, and imbalance-aware optimization. Recent studies on the role of artificial intelligence in cybersecurity have also highlighted the growing importance of machine and deep learning in automated security threat detection [5]. However, architectural complexity does not solve evaluation dependence on factors such as dataset version, class grouping and balancing, preprocessing sequence, and the relationship between training and test traffic.

A. Attention-Based Classification of Network Flows

The numeric properties of traffic flow data differ in importance across attack families. Attention captures interactions among these properties rather than weighting them equally. The self-attention mechanism was introduced in the Transformer model to establish relations between elements of the input sequence [6]. Zegarra Rodríguez et al. considered an attention-based approach to feature selection in IoT intrusion detection and obtained satisfactory results [7]. Tseng et al. investigated multi-class classification of the CIC-IoT-2023 dataset based on Transformer architecture [8]. Although these papers demonstrate the feasibility of attention-based attack-family classification, their performance measures cannot be directly compared with those obtained on the seven-group CICIDS2017 partition because of differences in traffic recording, attack definitions, class imbalance, and metrics. Therefore, this paper compares the Transformer model and other classifiers on the same CICIDS2017 partitions.

B. Minority Attacks and Representation Learning

Another limitation of multi-class intrusion detection is class imbalance, which means a classifier can be accurate when trained on the dataset but may behave less predictably on some minority attack families. Ullah et al. investigated Transfer Learning on Imbalanced Network Traffic using Transformers [9]. This line of research shows that building models under class imbalance is worth studying. However, class imbalance is a completely different problem from split leakage, so solving one does not prove the absence of the other. Thus, the proposed experiment requires testing both the model's performance on the minority classes and the integrity of the partition. Autoencoders can also enhance input data for the classifier by learning representations through reconstruction. These representations may reflect nonlinear interconnections among flow attributes. Researchers tested neural detectors based on such representations to detect and mitigate threats in enterprises [10]. Nevertheless, the reconstruction procedure may not capture all features that distinguish similar attacks. Sarıkaya et al. investigated the robust autoencoder-based intrusion-detection model in adversarial conditions [11]. The problem studied in the paper is different from that discussed in the present article.

C. Autoencoder-Transformer Combinations

Hybrid approaches leverage representation learning and attention to overcome each method's inherent deficiencies. Kamal and Mashaly proposed an autoencoder-dense-Transformer architecture for anomaly-based intrusion detection [12]. Salehiyan et al. examined a Transformer-GAN-autoencoder approach on edge and industrial IoT

datasets [13]. These examples show that learned representations can be used with further classification, but the architectural combinations differ. However, the reported results cannot isolate the effect of sample-dependent gating compared with concatenating raw data and latent representations in the current seven-class classification problem. It is necessary to have four models: raw-only, latent-only, fixed-fusion, and gated, trained and tested on the same data samples. Zhang et al. proposed a Transformer-CNN-BiLSTM combination for IoT intrusion detection [14]. Multi-component architectures can capture different feature dependencies, but their analysis is complicated because performance improvements may not reveal which component caused them when an experiment eliminates or substitutes components under equal conditions. The current article requires such an approach. The same applies to the paper's evaluation procedure: an improvement after a pre-processing step does not necessarily mean it improves the classifier.

D. Evaluation Integrity as the Remaining Gap

The reference provided is the closest architectural comparator, describing a gated autoencoder–Transformer, five stratified random seeds, tree and neural baselines, component ablations, and per-class metrics on a balanced CICIDS2017 seven-group subset. Thus, this protocol can serve as a reference protocol because its results are available. Nevertheless, it does not provide duplicate counts or the model's performance on session- or time-disjoint holdout datasets for the seven classes. This difference matters because fitting a scaler and autoencoder on the training dataset is enough to prevent learned-preprocessing leakage. Still, a separate audit is needed to determine whether duplicates or related traffic appear across partitions. Consequently, the remaining research question is not whether models need higher accuracy. It is necessary to identify whether the model advantage and recognition of minority classes reported by the authors remain the same after evaluating the same task on increasingly independent and well-documented partitions. To address this gap, three measures are possible: the amount of overlap between partitions; differences in macro-F1 and class-specific recall across split protocols; and the contribution of gating and other model components in each protocol separately. The table in Figure 2 separates these research questions. The limitation column contains information about the absence of this evidence for each specific study, not a statement about their methodological invalidity.

Table 2: Critical comparison of related work and the testable gap addressed by this article.

Research direction	Study	Dataset or task as relevant here	Main method	Evidence contributed	Unresolved issue for this article
Attentive features	Zegarra Rodríguez et al. [7]	Multiple CIC intrusion benchmarks	Attention and feature selection	Supports attentive learning on tabular traffic	No matched seven-group leakage audit
Multi-class Transformer	Tseng et al. [8]	CIC-IoT-2023	Transformer classifier	Tests IoT attack-family classification	Different dataset and class mapping
Imbalance-aware learning	Ullah et al. [9]	Imbalanced network traffic	Transformer transfer learning	Addresses minority-sensitive classification	Does not measure cross-split flow overlap
Autoencoder robustness	Sarıkaya et al. [11]	Intrusion detection under adversarial conditions	Robust autoencoder	Examines robustness of learned representations	Adversarial robustness differs from split independence
Hybrid classification	Kamal and Mashaly [12]	Anomaly-based intrusion detection	Autoencoder–dense–Transformer	Integrates latent learning and attention	Does not isolate this specific raw–latent gate on matched splits

Hybrid classification	Salehiyan et al. [13]	Edge and industrial IoT datasets	Transformer–GAN–autoencoder	Evaluates a different hybrid design	Cross-dataset scores are not a direct CICIDS2017 baseline
Multi-component attention	Zhang et al. [14]	IoT intrusion detection	Transformer and CNN–BiLSTM	Explores complementary model components	Component attribution requires matched ablations
Closest comparator	Supplied reference material (supplied; unverified bibliographic record)	Balanced CICIDS2017 seven-group subset	Gated AE–Transformer	Five-seed baseline, ablation, and per-class reference results	Duplicate and group/time-disjoint results not supplied
Present article	Proposed audited seven-group evaluation	CICIDS2017 with documented split units	Gated AE–Transformer plus leakage audit	Specifies a protocol that tests model and partition effects together	Empirical findings require the new runs

Attention, autoencoders, balance correction, and gating are each potential methods of improving intrusion detection. According to Table 2, the above-cited papers support evaluating these approaches, but their independent results do not show whether a detector generalizes to other flows in the specific capture. The current paper makes an important contribution by presenting a controlled experiment that tests detector generalizability: it maintains a seven-class problem setup, verifies duplicate and capture connections, adapts pre-processing learning with training captures, and tests the same baseline and ablated model versions across possible splits.

III. METHODOLOGY

This section outlines the study's evaluation methodology for a gated autoencoder–Transformer architecture for seven-class network intrusion detection. This research's key question is whether the model's superiority reported in the literature still holds when controlling for duplicate flow records, capture relationships, and preprocessing. Thus, the partition design becomes another experimental condition. In each valid condition, all baseline and ablation models share the same data partitions and seven-class label mappings, and the final test dataset is not accessible during feature fitting, hyperparameter tuning, or checkpoint selection.

A. Research Design and Evaluation Conditions

Three evaluation conditions are described below. Condition S1 replicates the stratified random split used in the reference material. It can serve as the architectural baseline. Condition S2 leaves exact duplicates in one partition and reviews overlapping near-duplicates. Condition S3 leaves whole capture sessions or time intervals if the available metadata and seven-class labeling are feasible. Condition S3 cannot be called a seven-class evaluation if any class is missing in the held-out partition. For a valid comparison, each condition must preserve the seven-class label mapping, input feature design, architecture configuration, and main metric. The comparison should focus on performance changes due to evaluation independence, not on irrelevant published results.

Table 3: Evaluation conditions and the question answered by each condition.

Condition	Partition unit	Duplicate control	Main question
S1: Reference stratified	Individual labeled records	Measure overlap; reproduce reference setting	What does the reported style of split measure?
S2: Duplicate-group-disjoint	Groups of identical feature records	Exact duplicates confined to one partition; near duplicates audited	Does removing cross-split duplication change performance?

S3: Group/time-disjoint	Feasible capture session or period	Group and duplicate relationships confined to one partition	Does performance persist on more independent traffic?

Table 3 implements increasing independence. In S1, we use the same stratified reference split; in S2, no exact duplicate sets cross partitions, and we audit for near duplicates; and in S3, we withhold entire capture sessions or time sets where all seven classes are present. We use the same labeling, models, and metrics to isolate partitioning's effect on performance.



Figure 2. A leakage-proofed experimental pipeline for the three partition conditions.

Figure 2 translates the framework into an operational pipeline with five steps: auditing provenance and duplicates, constructing the selected partition condition, preprocessing and fitting using training data alone, training the autoencoder and gated classifier, and validation-based selection followed by a single locked-test evaluation. The leakage barrier operates on all transformations that have been learned, namely feature selection, imputation, scaling, class balancing, and representation learning; each of these transformations is separately refit within the training segment of each split.

B. Dataset Audit and Seven-Class Mapping

CICIDS2017 is the main baseline, divided into the categories Benign, Bot/Infiltration, Brute Force, DDoS, DoS, PortScan, and Web Attack. Before modeling, the experiment documents the source release and file hashes, original labels, categorized labels, capture files, timestamps, and available session IDs. The generated manifest enables reproduction of the class mapping and partitions. The reference data includes 78 possible numeric attributes, with the baseline reference implementation selecting 49 attributes. The restricted data set includes 13,992 instances: 2,000 per category except 1,992 instances of the Bot/Infiltration category. The provided reference data does not explain why the Bot/Infiltration category has eight fewer instances than the other categories; it reports this as given, without assuming a reason. These are reference-study conditions, not automatic tallies for a fresh download release. Audit the release data and the extraction program before deciding the composition of the new study's table. Notably, the Web Attacks count in a provided source table is incompatible with its claim of having a 2,000-record controlled sample. We will create a canonical duplicate key based on the canonical representation of the candidate flow features. A secondary key will lack a label so we can find duplicates with inconsistent labels. We will measure the number of near duplicates using a similarity metric decided before examining final testing error rates.

C. Leakage-Resilient Preprocessing and Partitioning

We mark non-finite values using a documented cleaning protocol. For each partition, we learn all imputation values, the low-variance feature filter, the scaler, and the class weights from the training records alone. We then apply the fitted objects as is to the validation and test datasets. Validation determines hyperparameters and checkpointing; validation and test records are not used during autoencoder training. In an additional oversampling experiment, it applies only to the training records. The reference split conforms to the specified reference material: 35% for training, 35% for validation, and 30% for testing, using seeds 11, 22, 33, 44, and 55. For S2, the partitioning uses information on duplicate group membership. For S3, we know the selected capture/session/time and the number of independent groups per class. Temporal partitioning is infeasible for a seven-class problem because it excludes all seven classes from the test set.

Let D_{tr}, $D_{val,\ and\ Dte\ denote\ the\ training,\ validation,}$ and test partitions, and let T_{tr} denote the preprocessing transformation fitted using D_{tr} alone. The transformed inputs are

$$X_{tr}^{*} = T_{tr}(X_{tr}), \quad X_{val}^{*} = T_{tr}(X_{val}), \quad X_{te}^{*} = T_{tr}(X_{te}) \tag{1}$$

Equation (1) is the governing leakage constraint: no separate scaler, filter, or imputer is fitted to the validation or test distribution.

D. Autoencoder Feature Learning

The standardized features of the training flow i can be expressed as xi ∈ ℝd. The encoder learns the latent features zi ∈ ℝk through Eθ, and the decoder Dφ reconstructs the data:

$$z_i = E_\theta(x_i),\ x^*_i = D_\varphi(z_i) \quad (2)$$

Parameters of the autoencoders are trained using mean squared reconstruction error minimized over all training flows:

$$L_{AE} = (1/N_{tr}) \Sigma_i \|x_i - \hat{x}_i\|_2^2 \quad (3)$$

The summation in (3) includes all training flows, and Ntr is the total number of training flows. We use the validation reconstruction loss to select checkpoints. For the classification task, we remove the decoder and freeze the best-performing encoder. In our baseline configuration, we set d = 49 and k = 64; because feature filtering is repeated in a new split, we must determine d for each run.

E. Sample-Dependent Gated Fusion

The vector x and the frozen autoencoder embedding z generated from (2) are concatenated as follows:

$$u_i = [x_i ; z_i] \in \mathbb{R}^{d+k} \quad (4)$$

Weights for each element of the fused vector are determined via a learned sigmoid gate:

$$g_i = \sigma(W_g u_i + b_g),\ v_i = g_i \odot u_i \quad (5)$$

In this case, Wg ∈ ℝ(d+k)×(d+k) is the weight matrix, bg is the bias vector, σ is the element-wise sigmoid function, and ⊙ is element-wise multiplication. The gate can thus assign different weights to raw and latent data across flows. In the reference coordinates, the fused vector vi has 113 elements and is used in the Transformer classifier explained in the next section.

We evaluate the effect of gating through two ablation experiments: fixed concatenation and no-gate settings.

F. Transformer Classifier and Training Goal

The gated raw-latent embedding vi of (5) is fed into a small Transformer encoder classifier [6]. In the reference setup, the 113-dimensional gated embedding is initially linearly projected to the model width of 128 dimensions; it is unclear whether such projection takes place only once for the entire embedding or is done independently for a set of feature tokens – tokenization ambiguity as discussed below. The resulting embedding passes through two encoder layers with four attention heads and a 256-dimensional feed-forward network. The encoder extracts interactions between input features, and the subsequent classification layer predicts the traffic

class the flow belongs to among seven possibilities. We first train the autoencoder, then freeze the encoder while training the gated classifier.

The classifier in the reference setup is trained using focal loss [15],

$$FL(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \quad (6)$$

where pt stands for the probability of the predicted true class, αt is the weight of that class computed using training data, and γ is the focusing parameter. In the reference case, γ is set to 2 to pay more attention to hard examples; the class weights αt are computed only on the training-data subset; see [15]. We use the validation data to choose the checkpoint, and the test data only after selection. One aspect is not specified in the available documentation and should be mentioned in the experimental results: how exactly the combined vector is passed to the encoder as input tokens. If the vector is treated as a single entity and used as one token, it cannot attend across different feature tokens. Therefore, the tokenization approach must be stated explicitly.

Table 4: Reference settings and planned component tests.

Item	Reference setting	Methodological check in this article
Optimizer	Adam; learning rate 0.001	Record exact optimizer configuration per run
Autoencoder	50 epochs; batch size 128	Fit on training flows only
Classifier	20 epochs; batch size 64	Select checkpoint using validation
Transformer	2 encoder blocks; 4 heads; feed-forward size 256	Record the implemented feature-token construction
Focal loss	$\gamma = 2$	Compare with gated model without focal loss
Raw features	49 retained in reference run	Refit filter within each new training partition
Latent features	64 in reference run	Compare raw-only and latent-only models
Gating	Sample-dependent sigmoid weights	Compare with fixed concatenation
Random variation	Seeds 11, 22, 33, 44, 55	Retain matched partitions and per-seed outputs

Table 4 contains the settings for the reference model and the testing plan for this study. For each data split, we train the autoencoder and feature filter on training data only, and select the classifier checkpoint based on validation data. We will use the raw-only, latent-only, and fixed-fusion models to evaluate the impact of the learned gate, and the gated model without focal loss to evaluate the impact of focal loss. We will use five seeds with matched splits to estimate run-to-run performance variation. We have yet to validate tokenization of features from the Transformer.

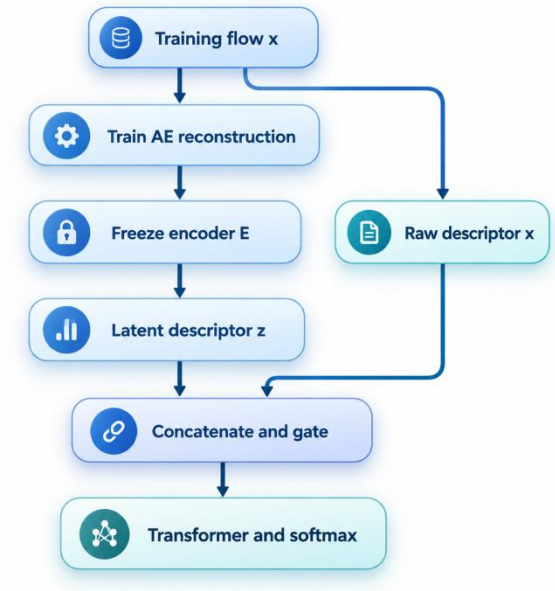


Figure 3. Raw-latent gated autoencoder-Transformer: Training and inference pipeline.

Figure 3 displays the pipeline utilized within each partition. The autoencoder is trained on the training partition. Its encoder is frozen to generate the latent feature z, and the standardized raw feature x is merged with z via the sample-specific gate before Transformer classification. This figure shows only the data flow, but the exact feature-to-token mapping construction is an implementation detail that needs documentation.

G. Performance Metrics and Statistical Analysis Plan

Macro-F1 score is the main performance metric because the main comparison treats each of the seven traffic classes equally. The secondary metrics include balanced accuracy, class-specific precision, recall, and F1, one-vs-rest precision-recall characteristics, benign false-positive rate, confusion matrices, and inference efficiency. All model comparisons will use the same partition indices and seed values. Results will include values by seed and mean ± standard deviation and, whenever possible, will compare paired models using confidence intervals estimated via bootstrapping. Multiple ablation comparisons will involve a multiplicity correction procedure. If any class is absent in S3, a seven-class macro evaluation will not be reported.

H. Reproducibility and Leakage-Safe Acceptance Criteria

To be deemed suitable for a leakage-safety claim, four criteria need to be met: no identical group should appear across partitions S2 and S3; all transformation models should only be fit on the training set; only the validation set should determine the hyperparameter and checkpoint choice; and the test set should be sealed until testing time. In the reproducibility package, we need to maintain hashes of the source files, label map, partition files, seeds, chosen features, preprocessing objects, model configuration (including Transformer tokenization), checkpoints, prediction files, and the hardware/software timing protocol.

IV. DATASET AND EXPERIMENTAL SETUP

This model relies on CICIDS2017 to classify seven network intrusion detection classes. The source materials provide a balanced set of references, including the features used, split ratio, training parameters, and metrics reported earlier. Our proposed evaluation will use reference conditions (S1), duplicate-aware conditions (S2), and feasible group/time-disjointed conditions (S3). In this part, we separate the parameters described in the source materials from those required to be generated later.

A. Dataset Origin and Class Distribution

The CICIDS2017 dataset consists of labeled network flow records collected under varying traffic and attacks [1]. The provided reference information categorizes its labels as Benign, Bot/Infiltration, Brute Force, DDoS, DoS, PortScan, and Web Attacks. To enable reproducibility, the dataset record will include details on its origin and release, filenames, file hash values, extraction date, the mapping from original to grouped labels, and the number of records retained after each cleaning process. The capture date and any session identifiers will be recorded in an audit document, regardless of whether they are included in model training. The provided reference information shows a 13,992-record control dataset, whose distribution is given in Table 5 below.

Table 5: Seven-class controlled sample reported in the supplied reference material.

Grouped class	Records in controlled subset	Reported test support per seed
Benign	2,000	600
Bot/Infiltration	1,992	598
Brute Force	2,000	600
DDoS	2,000	600
DoS	2,000	600
PortScan	2,000	600
Web Attack	2,000	600

Total	13,992	4,198

According to Table 5, the reference comparison took place under conditions of approximately equal class support. This helps with component comparison in the model, but it leads to different prior classes in traffic than those used operationally. During dataset creation, one test is needed: another table in the provided information shows 218 Web Attack examples in the pre-balanced set, whereas the controlled set includes 2,000 Web Attack examples.

B. Features, Missing Values, and Categorical Variables

The provided reference material states there are 78 numerical flow descriptors as potential candidates, of which only 49 are kept using its feature selection technique. Include the names of the retained features in the reproducibility pack for the new experiment. Because the filter is trained independently for each training set, the number of retained features is not fixed at 49; report it separately for each training-set split. The reference approach uses its 49 numerical features as an input to an autoencoder embedding model that produces 64-dimensional latent variables, resulting in a concatenated vector of 113 dimensions – raw and latent. Before training any model, validate and clean the dataset for missing fields, positive or negative infinities, invalid numeric values, constant columns, duplicated rows, and identification columns that describe the capture scenario. Each data-cleaning step should report the number of affected rows/columns. For imputing missing numeric values, compute them using the training records and apply them unchanged to the validation and test sets. The reference approach describes a dataset after cleaning its numerical features; it does not guarantee encoding of categorical predictors. If the new setup involves categorical input, build the vocabulary from the training data and define a pre-defined policy to handle unseen categories in the validation/testing datasets.

C. Splitting and Leakage Prevention

The reference materials follow a stratified split of 35% training, 35% validation, and 30% test, with seeds 11, 22, 33, 44, and 55. In each split, we create the data partition before applying the feature filter, imputer (if any), scaler, autoencoder, class weights, or classifier. We perform validation tuning and set aside test-set observations for evaluation only. Table 3 presents the three partition conditions considered in the leakage analysis. Perform exact feature hashing and label-independent duplicate detection before partitioning the data. Set the near-duplicate condition before evaluating the final test-set error. For S3, state the number of independent partitions and the class support in each partition. If S3's held-out test set does not contain all seven classes, S3 cannot be considered a seven-class comparison.

D. Scaling and Imbalance Handling Procedure

We estimate scaling parameters using only training samples. We then apply the same transformation to the validation and test sets. We train autoencoders on the training dataset; after selecting a checkpoint, we feed the encoder's frozen latent features into the classifier. Focal loss class weights are estimated using training samples, not validation and test class proportions. The reference subset has been intentionally balanced to allow controlled comparison. The new experiment should report counts for each class in the original dataset, in the clean dataset (after any sampling performed exclusively during training), and in the untouched test set. For oversampling evaluation, it will be done only within the training dataset. An unbiased test set, when supported by the source extraction, must be reported separately from the balanced diagnosis subset so that precision and false alarms are assessed under the correct class priors.

E. Hardware and Software Information

Table 4 provides information on the reference optimizer, number of epochs, Transformer configuration, focal loss application, feature dimensions, and seeds. Therefore, Table 6 focuses on reproducibility information not inferable from the provided results: the encoder and embedding dimensions used, the hardware (CPU/GPU and RAM), software and library versions, and the timing protocol (batch size, memory usage).

Table 6: Required Reproducibility and Execution Records

Component	Setting reported in the supplied reference material	Required run record

Latent representation	64 dimensions for the seven-class reference task	Encoder weights and embedding dimensions
Hardware and software	Not specified in the supplied result tables	Processor or GPU, RAM, framework and package versions
Efficiency	Per-sample inference time reported (ms/sample)	Batch size, timing method, memory, and throughput

As noted in Table 4, the implementation should also indicate the tokenization process for each feature, as the architecture diagram and description allow for multiple such processes.

V. RESULTS AND COMPARATIVE ANALYSIS

The following summarizes data from the provided reference literature to prepare the planned comparison. These values are not results obtained by the leakage-safe framework developed in Sections III and IV. We did not run conditions S2 and S3; therefore, we cannot present novel performance metrics or deployment parameters. Bibliographic information and restrictions regarding the provided literature are discussed explicitly in Section VI.

A. Primary Reported Results and Baselines

In the reference study, the proposed gated autoencoder-Transformer achieved 99.57 ± 0.04% accuracy and 99.61 ± 0.05% macro-F1, respectively, over five seeds. The macro-F1 value is 0.32 percentage points higher than XGBoost, but it requires about four times the per-sample inference time. Table 7 shows the three reported baselines.

Table 7: Previously reported results on balanced CICIDS2017 seven-class subset

Model	Accuracy (%)	Macro-F1 (%)	Inference (ms/sample)
Random Forest	99.22 ± 0.16	99.22 ± 0.16	0.034 ± 0.040
XGBoost	99.29 ± 0.16	99.29 ± 0.16	0.032 ± 0.037
Gated AE–Transformer	99.57 ± 0.04	99.61 ± 0.05	0.132 ± 0.009

The margin in Table 7 over tree-based baselines is small in absolute terms and comes with a higher inference cost. Because we evaluated all three models using a single stratified procedure, the table cannot reflect any change in that margin under duplicate-aware or group-disjoint splits.

B. Reported Ablation and Per-Class Performance

The ablation shows that combining raw and autoencoded features performed better than using each feature individually. From Figure 4, the raw-feature MLP scored close to 98.79% in macro-F1, while fixed raw–latent fusion scored around 99.01%; the provided document does not include a table with these two numbers, so we report them here approximately based on Figure 4. However, we can tabulate the gated variants precisely: gating without focal loss scored 99.21 ± 0.11% in macro-F1, while the full model scored 99.61 ± 0.05%.

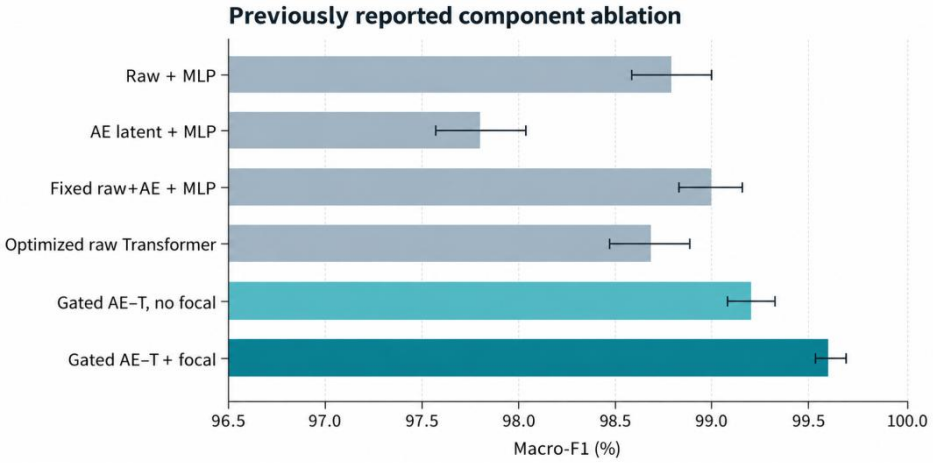


Figure 4: Component ablations found in the provided reference materials.

In Figure 4, error bars represent standard deviations based on five seeds, and the x-axis is limited to 96.5% to be able to see small variations. The first reading from the bottom up, on the weakest variation, is that of the autoencoder latent MLP, followed by the raw-only Transformer and the raw feature MLP; raw-latent concatenation outperforms all three, while the two gated variations outperform all four. The ranking is thus consistent with the assumption that the gate provides an advantage over simple concatenation. The difference between the performance of the complete model and its closest variation (without focal loss gate: 99.61% vs 99.21%) is several times larger than either variation's standard deviation. Still, the authors provide no per-seed comparison; thus, they describe the difference but do not show it is statistically significant. All seven classes achieve a reported F1 of at least 99.30%. Bot/Infiltration had the lowest recall (99.18 ± 0.15%) and the lowest F1 (99.30 ± 0.12%). Table 8 provides detailed information for each class, not just accuracy.

Table 8: Reported per-class recall and F1 of the gated AE-Transformer

Class	Recall (%)	F1 (%)
Benign	99.83 ± 0.05	99.79 ± 0.05
Bot/Infiltration	99.18 ± 0.15	99.30 ± 0.12
Brute Force	99.64 ± 0.07	99.68 ± 0.05
DDoS	99.55 ± 0.08	99.61 ± 0.06
DoS	99.86 ± 0.04	99.83 ± 0.04
PortScan	99.51 ± 0.09	99.57 ± 0.07
Web Attack	99.42 ± 0.12	99.50 ± 0.10

Across all seven traffic classes, recall and F1 scores were above 99%. The DoS class had the highest recall (99.86%). The Benign and DoS classes had the highest F1 scores of 99.83%, and the Bot/Infiltration class had the lowest recall, making it the hardest class to detect.

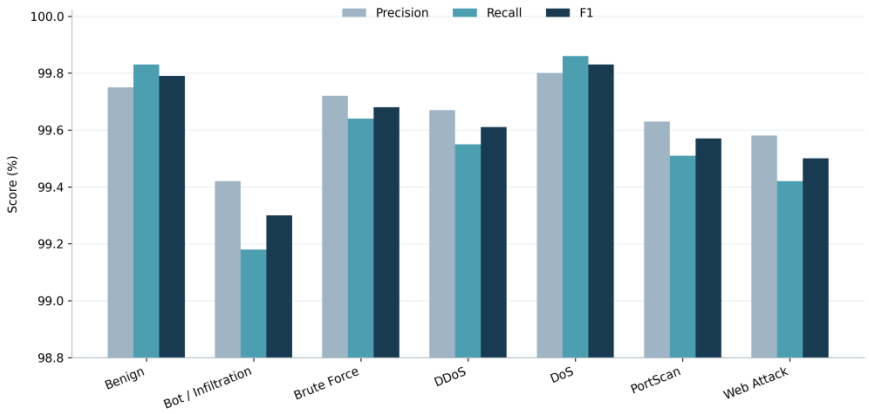


Figure 5. Precision, recall, and F1 for all seven CICIDS2017 classes in the provided reference material.

As shown in Figure 5, precision, recall, and F1 values are very high for all seven CICIDS2017 classes. However, although this is a balanced test set, the Bot/Infiltration class appears to be the weakest. That is why this class deserves attention in further research on the model's performance in duplicate-aware and group-disjoint splits. The vertical axis starts at 98.8%, so the differences look more significant. Table 8 contains recall and F1 values for each class, but not precision.

C. Requirements for Leakage-Safe Validation

Leakage-safe validation is deemed complete once the empirical package includes: (i) exact and near-duplicate audit counts for each split configuration; (ii) class support tables that check if each withheld partition supports all seven classes; (iii) baseline and ablated output at the seed level; (iv) full confusion matrices and precision-recall analysis based on probabilities from locked-test predictions; and (v) timing results from end-to-end runs on stated hardware. S2 and S3 are not intended to require performance improvements beyond S1; rather, they check whether the

architectural claims hold up under greater traffic independence. Thus, a drop in performance under more restrictive splits is meaningful, not an experiment gone wrong.

D. Discussion

The reference reports a macro-F1 of 99.61% on a balanced seven-class CICIDS2017 subset. In their ablation sequence, the authors ranked the full gated autoencoder-Transformer above the raw features, latent features, fixed fusion, and no-focal-loss versions. This may be because sample-specific gating helps the classifier modulate its reliance on both raw and latent features. However, this is an interpretative conclusion without statistical significance, since we did not provide seed-level predictions. The reference results are encouraging because they classify the seven attack groups in a controlled environment, but this does not imply equivalent performance under operational conditions. "Bot/Infiltration" has the lowest recall rate (99.18%) and should therefore be the focus of future rigorous experiments. In line with cyber threat-intelligence standards [16], a security operation center should also measure false positives under natural class-distribution conditions; the modern SOC and MITRE ATT&CK communities describe the relevant methodology [17]. More extensive research on AI in cybersecurity also highlights the distinction between benchmark performance and operational utility [18]. The reported inference time of 0.132 ms per sample is only a reference value because the authors do not document preprocessing steps, traffic collection, batching, hardware, memory, or latency.

VI. LIMITATIONS AND THREATS TO VALIDITY

First, the reference experiment is based on a balanced subset of 13,992 samples, which is unrepresentative of attack frequencies in a production network and, in addition, does not demonstrate performance under the duplicate-aware S2 or group/time-disjoint S3 settings because it uses a stratified train/validation/test split. Flows that may belong to each other, or even repeated records, may make random splitting simpler than with new traffic. Thus, in this paper, leakage-safe refers to the evaluation framework design rather than proven S2/S3 performance, which requires the specified audits and training-only transformations with independent holdouts. CICIDS2017 alone is not enough to show generalization to modern IoT devices, other networks, concept drift, or unseen advanced persistent threats, and grouping Bot/Infiltration labels is not a valid measurement. Another limitation related to source integrity is the following: the provided source has 99.61% macro-F1 in its tables but 99.64% in the concluding passage, and neither the author, conference/journal, year, volume, issue, page number, nor DOI could be found. The article thus uses 99.61% as a reference value, not as peer-reviewed evidence. The paper provides no figures on significance, ROC/AUC, or a seven-class confusion matrix because the predictions are not available.

VII. CONCLUSION AND FUTURE WORK

This work provides a leakage-safe evaluation protocol for gated autoencoder–Transformer-based multi-class network intrusion detection [19]. Rather than focusing only on high accuracy on benchmark datasets, the protocol accounts for duplicate contamination, partition independence, capture relations, and training-only preprocessing as necessary aspects of evaluation [20]. The S1-S3 protocol gradually moves from conventional stratified reference partitioning to duplicate-group-disjoint and increasingly independent group/time-disjoint evaluation while keeping class mappings, model architectures, and evaluation metrics intact. This protocol also includes raw and latent feature combinations using sample-specific gating and matches baselines, ablations, class-specific metrics, statistical methods, and reproducibility criteria [21]. The previously reported Macro-F1 score of 99.61 ± 0.05% is provided as a reference value, not as a leakage-safe result of this study. Future work will implement the full S1, S2, and S3 experimental protocol on the CICIDS2017 source data with verified partitions, paying special attention to exact and near-duplicate evaluation, class-support constraints, and the robustness of the raw-latent gated combination under increasing traffic independence. This implementation will also document the Transformer feature-to-token conversion process and test raw-only, latent-only, fixed-fusion, no-focal-loss, and complete-model variants with matched seeds and partitions. Future experiments should report class-specific precision, recall, and F1-scores; balanced accuracy; false positive rates; confusion matrices; precision-recall performance; confidence-interval comparisons; computational cost; memory consumption; and inference speed [22]. Validation on natural

class distributions and external, independent IoT or network security datasets is also needed to assess robustness to distribution shifts, concept drift, or unseen attacks before claiming operational generalization [23].

REFERENCES:

[1] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," in *Proceedings of the 4th International Conference on Information Systems Security and Privacy*, 2018.

[2] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why Do Tree-Based Models Still Outperform Deep Learning on Tabular Data?" *Advances in Neural Information Processing Systems*, vol. 35, 2022.

[3] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in Data Mining: Formulation, Detection, and Avoidance," *ACM Transactions on Knowledge Discovery from Data*, vol. 6, no. 4, 2012, doi: 10.1145/2382577.2382579.

[4] F. Pendlebury, F. Pierazzi, R. Jordaney, J. Kinder, and L. Cavallaro, "TESSERACT: Eliminating Experimental Bias in Malware Classification across Space and Time," in *28th USENIX Security Symposium*, 2019.

[5] A. Ali, A. Razzaque, U. Munir, et al., "AI-Driven Approaches to Cybersecurity: The Impact of Machine and Deep Learning," in *2024 2nd International Conference on Cyber Resilience (ICCR)*, IEEE, 2024.

[6] A. Vaswani, N. Shazeer, N. Parmar, et al., "Attention Is All You Need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[7] D. Zegarra Rodríguez et al., "Attentive Transformer Deep Learning Algorithm for Intrusion Detection on IoT Systems," *PLOS ONE*, vol. 18, no. 10, e0286652, 2023, doi: 10.1371/journal.pone.0286652.

[8] S.-M. Tseng et al., "Multi-Class Intrusion Detection Based on Transformer for Internet of Things," *Future Internet*, vol. 16, no. 8, p. 284, 2024, doi: 10.3390/fi16080284.

[9] F. Ullah, S. Ullah, G. Srivastava, and J. C.-W. Lin, "IDS-INT: Intrusion Detection System Using Transformer-Based Transfer Learning for Imbalanced Network Traffic," *Digital Communications and Networks*, vol. 10, no. 1, pp. 190–204, 2024, doi: 10.1016/j.dcan.2023.03.008.

[10] A. Ali, A. Zia, A. Razzaque, et al., "Enhancing Cybersecurity with Artificial Neural Networks: A Study on Threat Detection and Mitigation Strategies," in *2024 2nd International Conference on Cyber Resilience (ICCR)*, IEEE, 2024.

[11] A. Sarıkaya, B. G. Kılıç, and M. Demirci, "RAIDS: Robust Autoencoder-Based Intrusion Detection System Model against Adversarial Attacks," *Computers & Security*, vol. 135, p. 103483, 2023, doi: 10.1016/j.cose.2023.103483.

[12] H. Kamal and M. Mashaly, "AE-DTNN: Autoencoder–Dense–Transformer Neural Network Model for Efficient Anomaly-Based Intrusion Detection Systems," *Machine Learning and Knowledge Extraction*, vol. 7, no. 3, p. 78, 2025, doi: 10.3390/make7030078.

[13] A. Salehiyan, P. S. Moghaddam, and M. Kaveh, "An Optimized Transformer–GAN–AE for Intrusion Detection in Edge and IIoT Systems: Experimental Insights from WUSTL-IIoT-2021, EdgeIIoTset, and TON_IoT Datasets," *Future Internet*, vol. 17, no. 7, p. 279, 2025, doi: 10.3390/fi17070279.

[14] C. Zhang, J. Li, N. Wang, and D. Zhang, "Research on Intrusion Detection Method Based on Transformer and CNN-BiLSTM in Internet of Things," *Sensors*, vol. 25, no. 9, p. 2725, 2025, doi: 10.3390/s25092725.

[15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[16] A. Ali and B. M. Bhatti, *Spies in the Bits and Bytes: The Art of Cyber Threat Intelligence*, 1st ed. CRC Press, 2024, doi: 10.1201/9781003504108.

[17] S. G. Virk, J. Iqbal, A. Ali, A. R. Mahmud, I. Rashid, and T. Hanif, "Advancing Security Operations Centers: Modern Use Cases, MITRE ATT&CK Integration, and Coverage Optimization in 2025," *Journal of Computing & Biomedical Informatics*, vol. 9, no. 02, 2025.

[18] A. Ali, S. Ahmed, M. Hussain, et al., "The Synergy of Artificial Intelligence and Cybersecurity," in *2024 2nd International Conference on Cyber Resilience (ICCR)*, IEEE, 2024.

[19] A. R. Mahmud, A. Ali, M. R. Ajmal, S. G. Virk, S. Sheikh, and M. T. Khan, "A Federated Intrusion Detection Approach for SOC-Centered APT Defense Systems," in *Proceedings of the 11th International Conference on Next Generation Computing (ICNGC)*, 2025, pp. 193–196.

[20] A. Ali, M. A. Khan, K. Farid, S. S. Akbar, A. Ilyas, T. M. Ghazal, and H. Al Hamadi, "The Effect of Artificial Intelligence on Cybersecurity," in *2023 International Conference on Business Analytics for Technology and Security (ICBATS)*, 2023, doi: 10.1109/ICBATS57792.2023.10111151.

[21] N. A. Al-Dmour, M. Kamrul Hasan, M. Ajmal, M. Ali, I. Naseer, A. Ali, H. A. Hamadi, and N. Ali, "An Automated Platform for Gathering and Managing Open-Source Cyber Threat Intelligence," in *2023 International Conference on Business Analytics for Technology and Security (ICBATS)*, 2023, doi: 10.1109/ICBATS57792.2023.10111470.

[22] S. A. Changazi, I. Shafi, K. Saleh, M. H. Islam, S. M. Hussainn, and A. Ali, "Performance Enhancement of Snort IDS Through Kernel Modification," in *2019 8th International Conference on Information and Communication Technologies (ICICT)*, pp. 155–161, 2019.

[23] S. M. Hussain, M. H. Islam, A. Ali, and M. U. Nazir, "Threat Modeling Framework for Security of Unified Storages in Private Data Centers," in *2020 IEEE 22nd Conference on Business Informatics (CBI)*, vol. 2, pp. 111–120, 2020.

Declarations

Funding. No external funding was received.

Conflicts of interest. No conflicts of interest.

Ethics and consent. Not applicable statement.

Data availability. Not applicable.

Code availability. Not applicable.

Author contributions. CRediT roles and initials.

Acknowledgments. Optional.

Use of generative AI. Disclose AI tool use, or state that none was used.

How to Cite This Article

SUGGESTED CITATION

A. Javed et al., "A Leakage-Safe Evaluation Framework for Gated Autoencoder–Transformer Multi-Class Network Intrusion Detection," International Journal of Security, Intelligence, Governance & Modern AI (IJSIGMA), vol. 1, no. 1, pp. 34–50, 2026, doi: [DOI to be assigned].

ijsigma.org

SIGMA International Journal of Security, Intelligence, Governance & Modern AI

VOLUME 01 · ISSUE 01 · 2026 · ARTICLE 0003

RESEARCH ARTICLE OPEN ACCESS · ISSN: [ISSN to be assigned] · DOI: [DOI to be assigned]

AI-Generated Misinformation and the Psychology of Belief Formation: Investigating Cognitive Biases, Emotional Manipulation, and Susceptibility to Digital Radicalization

Saba Nadeem1,* and Almas Irfan1

1 Department of Psychology, Rawalpindi Women University

Corresponding author: Saba Nadeem (sabahkhan1130@gmail.com) · ORCID: [0000-0000-0000-0000]

Author emails: Almas Irfan (almas.irfan@f.rwu.edu.pk)

ABSTRACT

Costs involved in the generation of fluent written material, imagery, and audiovisual manipulation have been reduced with the advent of Generative AI. This creates a potential risk for belief formation and diffusion. While cognitive biases, emotional manipulation, and online radicalization have all been studied individually, there is little knowledge about the relationship between exposure to content, belief formation, susceptibility to emotional manipulation, and radicalization through AI-generated material. This paper develops a theory-informed conceptual framework explaining the process through which cognitive biases and emotional manipulation affect perception of reality and belief formation, intentions to share, and attitude susceptibility towards radicalization. It is a conceptual framework paper supported by a preregistration protocol for a randomized online experiment. The theoretical framework includes dual-process theory, motivated reasoning, illusory truth effect, social identity theory, and emotional appraisal integrated into a moderated mediation model. The framework yields five testable hypotheses and treats exposure, belief, endorsement, and participation as different dependent variables.

Keywords: AI-generated misinformation; belief formation; cognitive biases; emotional manipulation; generative artificial intelligence; digital radicalization; attitudinal susceptibility; media literacy

Highlights

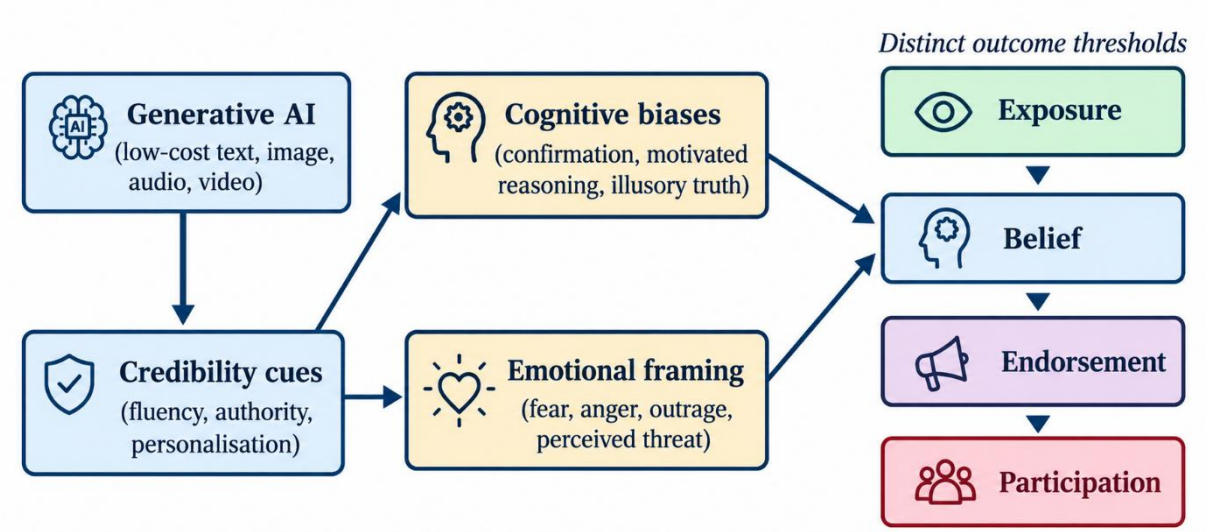
- Costs involved in the generation of fluent written material, imagery, and…
- This creates a potential risk for belief formation and diffusion.
- While cognitive biases, emotional manipulation, and online radicalization have…

IMPACT STATEMENT

Costs involved in the generation of fluent written material, imagery, and audiovisual manipulation have been reduced with the advent of Generative AI. This creates a potential risk for belief formation and diffusion. While cognitive biases, emotional manipulation, and online radicalization have all been studied individually, there is little knowledge about the relationship between exposure to content, belief formation, susceptibility to emotional manipulation, and radicalization through AI-generated material. This paper develops a theory-informed conceptual framework explaining the process through which cognitive biases and emotional manipulation affect perception of reality and belief formation, intentions to share, and attitude susceptibility towards radicalization. It is a conceptual framework paper supported by a preregistration protocol for a randomized online experiment. The theoretical framework includes dual-process theory, motivated reasoning, illusory truth effect, social identity theory, and emotional appraisal integrated into a moderated mediation model. The framework yields five testable hypotheses and treats exposure, belief, endorsement, and…

I. INTRODUCTION

Generative artificial intelligence technology enables the production of persuasive written text, imagery, audio, and video to be inexpensive and rapid [1]; yet, it presents a psychological problem that cannot be addressed through technical identification alone: in what conditions do people end up believing, spreading, and acting on this information? These issues will be addressed by this paper in the following five stages. First, the paper will define AI-created misinformation and differentiate it from disinformation. Second, it will describe the process of belief formation and updating. Third, it will identify a particular gap in the literature regarding cognitive bias, emotional manipulation, and vulnerability to radicalization. Fourth, it will outline the research objective and objectives.



1.1 Figure 1. Overview of the problem space, from AI-generated content to distinct outcome thresholds.

As shown in Figure 1, generative AI influences human beliefs and behavior through credibility cues, cognitive biases, and emotional framing, potentially leading to endorsement and participation. However, exposure to AI-generated content does not necessarily result in belief or active participation.

A. Background and Context

The abilities of generative AI systems to generate fluent text, photorealistic images, synthetic voices, and fake videos inexpensively, as well as customize messages according to the language, tone, and preferences of the target audience, make it easier to create misleading content and undermine the usefulness of traditional authenticity cues [2]. This paper will examine the psychological processes involved in the consumption of such content. Specifically, it will analyze how and to what extent biases and emotions influence the

evaluation of AI-generated misleading information. However, in this paper, we will not assume that AI-generated information is always misleading or that its consumption leads to beliefs and further actions.

B. Psychology of Belief Formation

The process of evaluating information involves both rapid intuitive and slow effortful processes. The cues of credibility include expertise, reputation, fluency, familiarity, social validation, and congruence with previous beliefs. In addition, they often do not undergo checks. Information that agrees with the beliefs of an individual and is delivered by a trusted in-group source is more likely to be accepted and analyzed less than other information. Repeating information makes it easier to process, which is why it may be taken for truth. Generative AI technologies can provide many of these cues inexpensively [5]; therefore, the process of reception of information will largely determine the impact of the technology.

C. Research Problem and Knowledge Gap

Earlier studies have considered three topics: (i) the cognitive psychology of accepting misinformation via the mechanisms of confirmation bias, motivated reasoning, and illusory truth effect; (ii) the role of emotions and emotional moral language in attracting attention and sharing; and (iii) social-psychological theories of radicalization. There are three gaps that require research. First, these three topics have not been merged into one framework that includes such outcomes as perceived truth, sharing, and susceptibility to radicalization. Second, the literature on AI-generated information mainly concerns detecting AI-generated content, while the interaction between synthetic credibility cues and cognitive/emotional factors requires more research [6]. Third, radicalization has mainly been viewed as an outcome of exposure to information, but not of endorsement. Emotional and cognitive biases have been known to influence the evaluation of misinformation; however, their interaction with regard to AI-generated content is unknown.

D. Research Objectives and Questions

The broad objective is to design an evidence-based theoretical model of the role that cognitive biases and emotional manipulation play in belief in AI-generated misinformation and susceptibility to digital radicalization. More specific objectives are: (O1) to estimate the effect of AI-generated misinformation on the perception of truth and belief; (O2) to estimate the effect of confirmation bias, motivated reasoning, and familiarity on the perception of credibility; (O3) to estimate the effect of emotional framing on processing and sharing intention; (O4) to assess the relationship between these factors and susceptibility to radicalizing narratives; and (O5) to design a model and test it. The following are four research questions to be answered in this study. RQ1: How strongly do cues of credibility and repetition predict perceptions of truth and beliefs? RQ2: Is there a moderating influence of belief congruence (which is used as a proxy for confirmation bias and motivated reasoning) on the predictive relationship? RQ3: What is the role of emotional framing in arousal, perception of credibility, and intention to share? RQ4: What is the role of arousal and identity-threatening situations in radicalization susceptibility besides beliefs?

E. Research Contributions

Contribution of the study: The contribution of the current paper is that it provides an empirically testable model that takes into account the combination of cognitive and emotional mechanisms in terms of their impact on three different types of outcomes (belief, sharing, and susceptibility). Additionally, the current paper provides recommendations on how to measure exposure and susceptibility, and not confuse one for the other.

II. LITERATURE REVIEW

However, the review will be more about synthesis and comparison of relevant literature rather than analysis of individual works. After presenting the basic types of AI-based misinformation and their perceptions of credibility and detectability, the review will move to discussing cognitive biases, framing, and social identity in terms of how they contribute to the acceptance of belief, attention to and forwarding of messages, and digital radicalization. In the course of discussion, established facts will be distinguished from hypotheses, and contradicting views will be identified.

A. AI-Generated Misinformation

Different forms of AI-generated misinformation exist. Generated synthesized texts may look credible because of fluency and authoritative style, and in one study, GPT-3-generated disinformation was evaluated to be more persuasive than human-written disinformation. Fake evidence (documents, citations, or statistics) may rely on credibility heuristics and anchoring; explainability-oriented language-model frameworks strive to increase the auditable nature of generated claims [7]. Deep fakes and manipulation of audio-video material depend on the credibility of recorded evidence and are especially believable in this sense. As for personalized narratives, they adjust to the interest and language of the in-group (what is technically achieved through user-profiling and tracking techniques [8]), and are expected to trigger confirmation and in-group biases, although the superiority over generic texts should not be taken for granted. The generation technologies are well-researched [9]; however, their psychological effects remain largely unknown. The human ability to detect generated material is weak: in one study, participants were unable to reliably differentiate between tweets written by GPT-3 and by real users.

B. Cognitive Biases and Belief Acceptance

Table 1 is devoted to mechanisms rather than synonyms. Confirmation bias deals with selective search and interpretation of evidence; motivated reasoning is goal-driven and identity-protective processing; the illusory truth effect is a fluency-based consequence of repetition that does not require any direction; availability and anchoring bias judgment because of easy recall and initial anchor points; authority and in-group biases act via source cues. Literature points to the strength of the repetition effect and congruence with existing beliefs, but differs with regard to the influence of analysis on identity-protective cognition and the length of time that changes in belief last.

Table 1. Cognitive biases and belief formation

Bias	Psychological meaning	Role with AI-generated content (illustrative example)	Boundary condition
Confirmation bias	Selective seeking and interpretation of evidence that supports existing views	A fluent AI-written article echoing a user's view is accepted without checking its claims.	Weaker when accuracy goals are salient or disconfirming evidence is easy to obtain
Motivated reasoning	Goal-directed processing that protects identity or preferred conclusions	Dismissing an accurate correction because it threatens group identity	Stronger for identity-relevant topics; weaker for low-salience issues
Illusory truth effect	Repeated statements feel truer because they are processed more fluently	An AI-generated claim recirculated across accounts feels familiar and therefore credible	Attenuated when prior knowledge is strong, or source is known to be unreliable
Availability heuristic	Judging frequency or risk by ease of recall	Vivid synthetic images of an event inflate perceived prevalence	Weaker when base-rate information is salient
Authority bias	Deference to perceived expertise	Fabricated expert quotation or professional formatting	Weaker when source is verifiably absent or flagged

Bias	Psychological meaning	Role with AI-generated content (illustrative example)	Boundary condition
In-group bias	Preferential trust in members of one's own group	Message adopting in-group idiom and symbols	Stronger under identity threat; weaker with cross-group norms of scrutiny

As can be seen from Table I, several psychological biases play an important role in perception and reaction to content created by AI. Confirmation bias, motivated reasoning, the illusory truth effect, availability heuristic, authority bias, and in-group bias could affect credibility perceptions and belief reinforcement. Yet, depending on different conditions (e.g., previous knowledge, personal relevance of the topic, trust in the source, access to counterevidence), their impact may vary.

C. Emotionally Manipulative Techniques

Fear, anger, anxiety, threat appeals, humiliation, and perceived injustice could attract attention, make people resort to heuristic cues, and stimulate sharing. In particular, moral-emotional language has been found to be associated with increased diffusion of political messages on social media, having higher impact in ideologically homogeneous networks than heterogeneous ones. Emotional persuasion is not a manipulative technique itself – manipulation is defined as a use of emotions for undermining reasoning or achieving deception. Emotions themselves may convey truthful information, and sharing, belief formation, and attitude change are not necessarily interconnected phenomena (people may share information without believing it). The question of whether AI-created emotional manipulation will be more efficient than human messaging is open, and the majority of the existing literature on the issue discusses technical attacks [10].

D. Social Identity and Digital Radicalization

In this case, digital radicalization is defined as a process of growing extremism of beliefs, attitudes, and behavior that promotes intergroup conflict at least partially through digital means [11]. Radicalization online does not necessarily mean exposure to radical content or endorsement of extremism and extremist acts, which are considered distinct phases of the process. Social identity, narratives of grievances, injustice perception, isolation, sense of belonging, and community endorsement can be considered potential influential factors. Group identity and threat theory, in particular, are especially significant among influential theories; yet the proportion of each factor, including exposure to algorithms, self-selection, and offline risk, still needs to be clarified [12]. There is no group specifically vulnerable to radicalization, and most people exposed to radical content are not radicalized.

E. Critical Synthesis and Research Gap

Previous established results: repetition and consistency with previous beliefs affect truth perception, and emotions affect attention and sharing. Previous contested topics: the general role of analytical thinking, the extent of identity protective effects, and the relationship between sharing and belief. Issues regarding research design: single-session studies, convenience sampling, laboratory artifacts, self-reports, and immediate results. Understudied topics: combined effect of credibility heuristics, repetition, and emotional frames on AI-produced information, and the effect of AI literacy. Prior AI-threat research has also focused more on cybersecurity and national defense [13], [14] rather than on individual beliefs. Inferences are hard when only mediators are measured, and susceptibility is difficult to measure due to low frequency and unethical experimental manipulation of radicalization, so attitude indicators are problematic.

F. Selected Theories

Dual process theory describes two types of processing modes intuitive and analytic. It explains the intuitive endorsement of fluency and authority, but it is criticized for being too simple a model and for failing to identify when analytic effort will occur. Both theories of motivated reasoning and confirmation bias posit a directional bias in processing, while directional bias may arise from prior-weighted updating as well. Hence, the concept of belief congruence serves as a moderating variable. The illusory truth effect predicts that repetition leads to perceived truth because of fluency, but its impact depends on the degree of prior knowledge. Social identity theory posits that group membership influences trust and attitudes towards other

groups, relating identity threat to the readiness to believe grievance narratives; however, it fails to explain why a very small percentage of people turn into extremists. The theories of emotional appraisal argue that threat, injustice, and loss lead to the emotional reaction that drives behavior and see motivation as an intermediate variable; however, manipulation and persuasion cannot be distinguished between them.

G. Proposed Conceptual Model

In order to bring together the psychological theories described above, the following conceptual model is presented as an attempt to explain the effect of generative AI on human behavior through cognitive biases, emotional framing, and perception of credibility [15]. This model separates exposure, belief, endorsement, and participation as distinct results of AI impact and does not treat exposure as something that will inevitably lead to the behavioral effect. In addition, it takes into account such boundary conditions as previous knowledge, identity relevance, reliability of the source, and goals of accuracy.

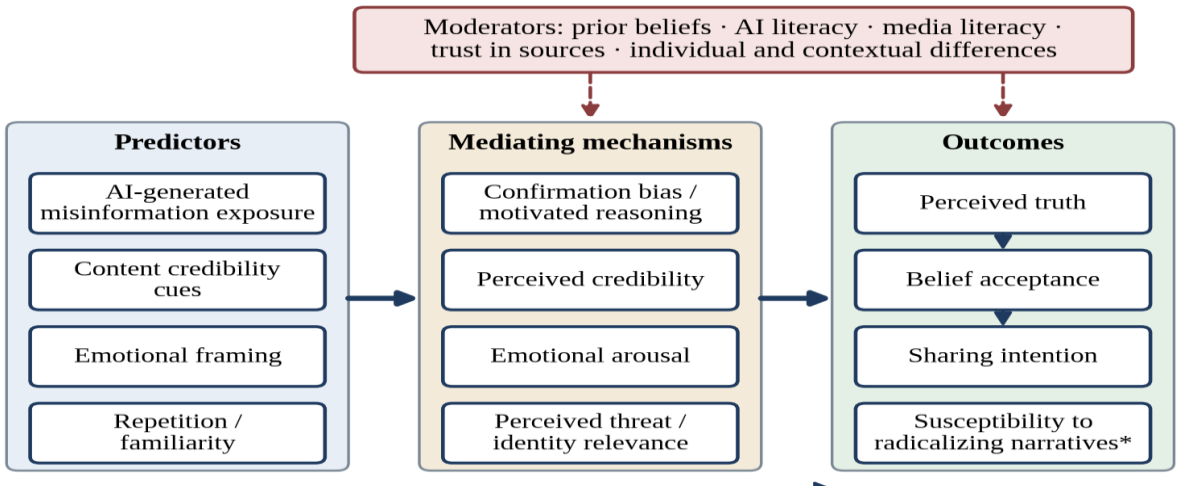


Figure 2. Conceptual framework of AI-generated misinformation and belief formation (hypothesized relationships).

The model is shown in Figure 2 below. Predictors (left side): Misinformation generated through AI technology, credibility cues, framing, repetition. Mediators (middle): Confirmation bias/motivated reasoning, perceived credibility, emotional arousal, threat/identity relevance. Outcome variables (right side): Perceived truthfulness, belief acceptance, sharing intentions, susceptibility to polarizing stories [16]. The solid lines represent the predicted relationships, whereas the dashed lines represent the predicted conditional relationships from the moderators (beliefs, media literacy, source credibility, situational factors). This figure captures the predicted relationships that will be empirically tested in the experiments as part of the hypothesis testing paths.

H. Research Hypotheses

The following research hypotheses are based on the proposed conceptual framework and theoretical concepts of cognitive psychology, morality, and AI-human interactions. These hypotheses explore the impact of AI-generated misinformation, perceived credibility, cognitive biases, and emotional framing on the process of belief formation, endorsement, and engagement [17, 18]. The hypotheses also take into account the moderating effect of individual and situational factors. Table 2 summarizes the five hypotheses along with justifications, variables, and analysis.

Table 2. Proposed hypotheses and evaluation methods

Hypothesis	Theoretical justification	Variables involved	Proposed analytical method
H1. Credibility cues increase belief acceptance, partly via perceived credibility.	Dual-process theory; authority and credibility heuristics	Credibility cues → perceived credibility → belief acceptance	Mediation with bootstrapped indirect effect (Eq. 2–3)
H2. Belief congruence	Motivated reasoning;	Exposure × belief	Moderation regression

Hypothesis	Theoretical justification	Variables involved	Proposed analytical method
strengthens the effect of exposure on perceived truth.	confirmation bias	congruence → perceived truth	with simple-slopes (Eq. 1)
H3. Emotional framing increases arousal, and arousal predicts sharing intention above and beyond belief.	Emotional appraisal; moral-emotional communication	Framing → arousal → sharing; belief as covariate	Mediation; hierarchical regression
H4. Repeated exposure increases perceived truth, less so for high-AI-literacy or high-prior-knowledge participants.	Illusory truth effect; fluency	Repetition × literacy / prior knowledge → perceived truth	Mixed-effects model; moderation
H5. Arousal and identity-relevant threat are positively associated with attitudinal susceptibility beyond belief acceptance.	Social identity theory; emotional appraisal	Arousal, threat, belief → susceptibility	Hierarchical regression; SEM if justified (association only)

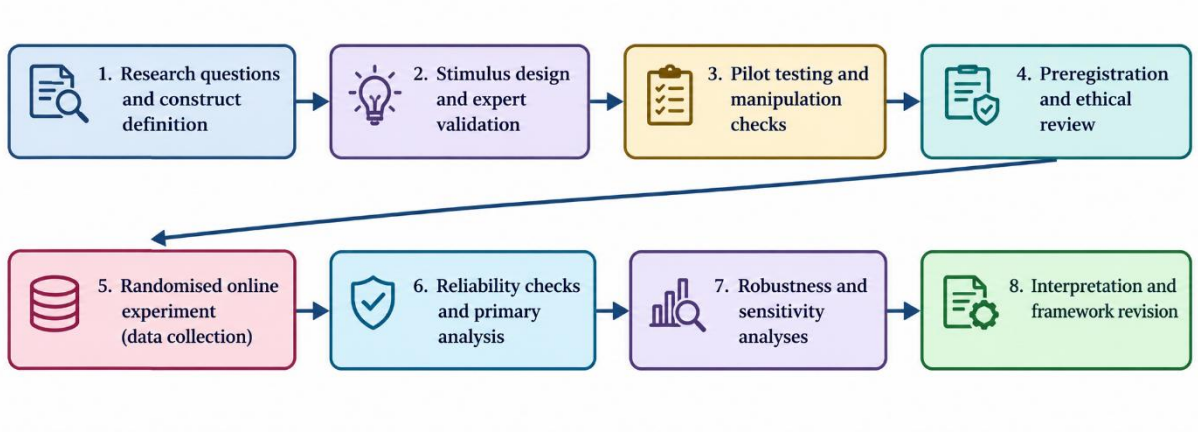
As shown in Table 2, the five research hypotheses explore the connections between the cues of credibility, congruence of beliefs, emotional framing, repetition, and susceptibility of individuals to AI-generated material. Each research hypothesis has a theoretical foundation in psychological theories and presents an appropriate way of analysis through mediation analysis, moderation regression, mixed-effects modelling, and structural equation modelling. The combination of these five research hypotheses provides the basis for an empirical examination of the suggested conceptual model.

III. METHODOLOGY

This section presents a reasonable and ethical approach to test the suggested framework. This includes a justification of the methodology and explanations of what each element is capable of showing: experimentally manipulated factors and relations between variables that are measured. In addition, the population, sampling, procedure, constructs, measurement, analyses, along with generic formulas for moderation and mediation will be outlined. Measures of validity, reliability, and safety for the participants will also be provided. No data have been collected yet.

A. Research Design

Two designs are used: conceptual framework development (Sections I–II), which formulates propositions through theoretical synthesis, and a proposed protocol for testing them, a preregistered randomized online experiment (not yet conducted). The experiment will manipulate emotional framing (neutral vs. high-arousal) and credibility cues (absent vs. present) between subjects, and repetition (single vs. repeated exposure) and message version (AI-generated vs. human-written) within subjects. Manipulated factors support experimental inference; measured mediators, moderators, and susceptibility support associations only (see Figure 3).



1.2 Figure 3. Proposed research workflow from construct definition to interpretation.

As illustrated in Figure 3, the research methodology follows an eight-step process to ensure systematic investigation, methodological rigor, and reliable findings. The process begins with defining research questions and constructs, followed by stimulus design, expert validation, pilot testing, and preregistration with ethical review. Subsequently, a randomized online experiment is conducted for data collection, followed by reliability assessment, primary analysis, and robustness checks. Finally, the findings are interpreted to refine the proposed conceptual framework and support the development of evidence-based conclusions.

B. Population, Sampling, and Procedure

Target population: adults (18+) who regularly use social media and are fluent in the language of the study; participants who fail the preregistered attention check are excluded. Recruitment will use online panels with quotas. The final sample size has not been set; it must be established before data collection by an a priori power analysis for the smallest effect size of interest (e.g., G*Power for ANOVA and regression; Monte Carlo simulation for mediation; $\alpha = .05$; power .80–.90; allowance for attrition). Procedure: consent; baseline measurement of prior beliefs and moderators (separate from stimulus exposure); random assignment; exposure to brief, matched, social-media-like messages on low-stakes disputed issues; measurement of outcomes and manipulation checks; and debriefing with corrections. Stimuli (AI-generated or human-written, by condition) will be checked by experts for equal length, readability, and factual status, and pretested for emotional intensity. No actual extremist recruitment material will be employed; the radicalization narratives will only be mild and fictionalized grievance-and-blame narratives.

C. Variables, Definitions, and Measures

The constructs are defined in Table 3 and distinguished between theoretical constructs and their operationalization's. For the variables for which there is no validated scale (AI literacy, perceived credibility, narrative susceptibility), the scale would need to be selected and adapted/validated prior to conducting the experiment (pre-test, factor analysis, validation). The self-report approach is vulnerable to the issue of social desirability and acquiescence; solutions to minimize these biases include anonymity, reverse keying, and attention checks. Susceptibility would be measured using attitudes towards the grievances and an appropriate measure of radicalism intention, yet to be selected.

Table 3. Research constructs and operationalization

Construct	Conceptual definition	Operational definition and measurement	Analytic role
Exposure; credibility cues; framing; repetition	Features of AI-generated content and its presentation	Manipulated conditions (AI vs. human version; cue present/absent; neutral vs. high-arousal; 1 vs. 3 exposures); manipulation checks	Independent variables
Confirmation bias /	Selective acceptance of belief-congruent	Belief congruence (baseline attitude × message direction); congruence-	Moderator/mediato r

Construct	Conceptual definition	Operational definition and measurement	Analytic role
motivated reasoning	information	weighted evaluation index	
Perceived credibility	Judged trustworthiness and accuracy of message and source	Multi-item rating scale (to be validated); reliability by α or ω	Mediator
Emotional arousal; threat/identity relevance	Activation and perceived danger to self or group	Self-reported arousal and discrete emotions; threat and identification items	Mediators
AI literacy; media literacy; source trust	Knowledge and dispositions that condition evaluation	Literacy tests (AI literacy requires development); trust scale	Moderators
Perceived truth; belief acceptance; sharing intention	Judged veracity; agreement held with confidence; stated intention to share	Truth rating; agreement × confidence; sharing-intention items (not behavior)	Outcomes
Susceptibility to radicalizing narratives	Attitudinal receptivity to grievance-based, us-versus-them messaging	Receptivity scale; validated intention instrument (to be selected). Not exposure or behavior	Outcome (association only)

As shown in Table 3 above, the research operationalizes the major constructs that are necessary in the investigation of how the AI-generated content impacts the judgment and behavioral intentions of human beings. In this case, the independent variables, mediators, moderators, and dependent variables have been identified and defined, along with a proposed method of measurement for each of them, to ensure a rigorous scientific investigation. These constructs include perceptions about the credibility of the message, emotional arousal, belief acceptance, intentions to share, and susceptibility to the radicalizing narrative.

D. Analytical Methods

Data analysis begins with descriptive statistics, testing for missing data using multiple imputation if data are missing randomly, conducting reliability tests (alpha or omega), and conducting manipulation checks. The effects of manipulations are measured using ANOVA or mixed models. The generic model for moderation analysis is

$$Yi = \beta 0 + \beta 1Xi + \beta 2Wi + \beta 3(Xi \times Wi) + \beta 4Ci + \varepsilon i \quad (1)$$

where Yi is the dependent variable of participant i, Xi is the predictor (e.g., exposure), Wi is the moderator (e.g., prior-belief congruence), Ci is the set of covariates, β is the coefficient vector to be estimated, and εi is the error term. Mediation is modeled as follows:

$$Mi = \alpha 0 + aXi + \alpha 2Ci + eMi \quad (2)$$

$$Yi = \gamma 0 + c'Xi + bMi + \gamma 2Ci + eYi \quad (3)$$

where Mi is the mediator, a is the effect of X on M, b is the effect of M on Y controlling for X, c′ is the direct effect, and a×b is the indirect effect, which will be assessed with bootstrapped confidence intervals. These are generic models to be specified in terms of the selected variables; no coefficients have been estimated. Because M is measured rather than manipulated, a×b can be interpreted causally only under additional assumptions (no unmeasured M–Y confounding, including confounders affected by X; correct temporal order; correct model specification); temporal ordering alone is not sufficient. Even where X is randomized, indirect effects will be reported as consistent with, not proof of, mediation. Structural equation modeling will

be employed only if measurement models fit well. Effect sizes with confidence intervals, Holm correction of p-values across confirmatory tests, and sensitivity analyses (alternative scoring, exclusion rules, and covariates) will be reported.

E. Reliability, Validity, and Ethics

Construct validity is supported by theory-driven operationalization, review of the stimuli by experts, and convergent and discriminant validation. Internal validity depends on random assignment and manipulation checks. External validity is constrained by the artificial nature of exposure and the sampling frame. Reliability is based on α or ω; selection bias is diagnosed through attrition analysis; reproducibility is supported by preregistration and open data/materials. Institutional review is mandatory before data collection (not yet obtained); consent includes the right to withdraw; data are minimized and anonymized; extremist recruitment content is avoided; participants receive signposted support; and debriefing refutes any false claims. Reporting is aggregate and non-stigmatizing.

IV. RESULTS

This section presents a conceptual synthesis and a proposed framework for evaluating the five hypotheses. Because the experiment has not been conducted, no participant data, statistical results, or simulations are reported, and nothing here should be read as an empirical finding. The section (i) traces the proposed pathway from exposure to belief updating, (ii) states, for each hypothesis, evaluation criteria that would support, qualify, or contradict it once data exist, (iii) discusses theoretical contributions and practical implications, and (iv) states limitations and future research.

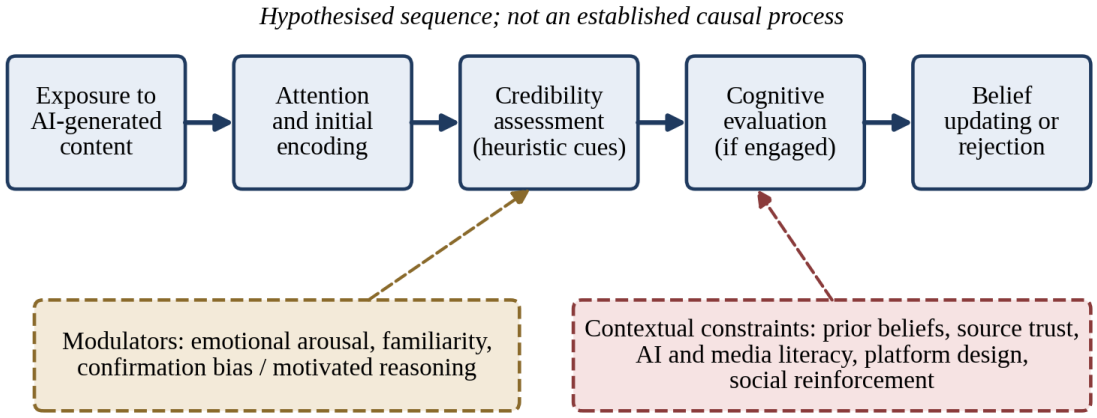


Figure 4. Proposed psychological pathway from exposure to belief updating (theoretical, not an established causal sequence).

The suggested process is depicted in Figure 4. Stages of this process will be affected by such factors as arousal, familiarity, and confirmation bias. Prior beliefs, trust towards the source, literacy, medium design, and social context will be the conditions for the stages. In this framework, exposure, perception of truth, belief acceptance, endorsement, sharing intention, susceptibility to radicalizing narratives, and radicalization itself or participation are regarded as different outcomes. Exposure does not imply perception of truth, sharing does not require belief, and susceptibility is not necessarily radicalization. It is believed that cognitive biases, arousal, credibility, repetition, prior beliefs, and social identity can interact with each other in the following way: repetition and credibility increase fluency and perceived credibility; prior beliefs and identity determine acceptance or rejection of such cues; and arousal, especially in case of identity threat, decreases scrutiny and increases sharing and acceptance of grievances. These relations are proposed and are not verified.

The criteria used for evaluation of hypotheses are provided in Figure 5. H1 (credibility cues ? perceived credibility ? belief acceptance) is supported by credible cue effect on belief and reliable indirect pathway through perceived credibility (it is conditioned by the measure of credibility), contradicting it in case of absence of this effect; H2 (exposure × belief congruence ? perceived truth) is supported by interaction when exposure increases the level of perceived truth in relation to congruent material, qualifying or contradicting it in absence of interaction or its restriction to identity-related content; H3 (framing ? arousal ? sharing, controlling for belief) is supported by a direct effect of arousal on sharing intention above belief acceptance,

qualifying it if both effects exist, contradicting it if the effect is absent; sharing intention is not sharing behavior. H4 (repetition × literacy or prior knowledge? perceived truth) is supported by the effect of repetition that is weakened with a higher level of literacy/knowledge, contradicting it in case of a constant effect; lack of validation of the literacy measure is a limitation. H5 (arousal and identity-relevant threat? attitudinal susceptibility beyond belief) is supported by association remaining after control for belief acceptance, contradicting it if it disappears; presence of association does not indicate causality.

Hypothesis	Would support	Would qualify	Would contradict
H1	Cues raise belief via credibility	Cue effect with no credibility mediation	No cue effect on belief
H2	Reliable exposure × congruence interaction	Interaction only for identity-relevant topics	No interaction (accuracy motive favoured)
H3	Arousal predicts sharing beyond belief	Arousal predicts both sharing and belief	Arousal unrelated to sharing
H4	Repetition raises truth; attenuated by literacy	Repetition effect equal across literacy levels	No repetition effect
H5	Arousal and threat linked to susceptibility	Link vanishes after controlling for belief	No association (not causal in any case)

Figure 5. Conceptual evaluation map (not empirical results) of result patterns that would support, qualify, or contradict each hypothesis.

Figure 5 is conceptual and contains no data. It substitutes possible patterns that would support, modify, or refute each hypothesis in lieu of illustrative interaction graphs that could be confused with empirical findings to ensure the framework remains falsifiable.

A. Possible Susceptibility Mechanisms and Alternative Explanations

It may be that an increased level of arousal resulting from the identity threat would make one more susceptible to believing the narrative of grievances. H5 will test only an association within one session. While an observed association between arousal, identity threat, and attitudinal susceptibility would not prove that misinformation leads to radicalization, attitudinal susceptibility is neither radicalization nor engagement in it. Possible alternative explanations include pre-existing beliefs, demand effects, overlap of measures, and omitted variable bias. Absence of difference between AI-generated and human-written messages in terms of susceptibility would indicate the importance of psychological processes, not the origin of information.

B. Interpretation and Theoretical Contributions

The process of acceptance of misinformation by AI is not governed by any one mechanism. Heuristic fluency, repetition, and authoritative signals work largely on a heuristic level, while congruence and identity influence the evaluative goals; emotions channel attention and motivation towards sharing, which is independent of belief formation. By bringing together both cognitive and emotional processes and making distinctions between outcome levels, the model perceives generative AI as a provider of credibility cues rather than an agent of psychology. This model also admits the possibility that congruence is a result of prior weight updating through rational means, thereby negating any possibility of a theory solely based on motivated reasoning. In the context of radicalization, this model differentiates between susceptibility and radicalization.

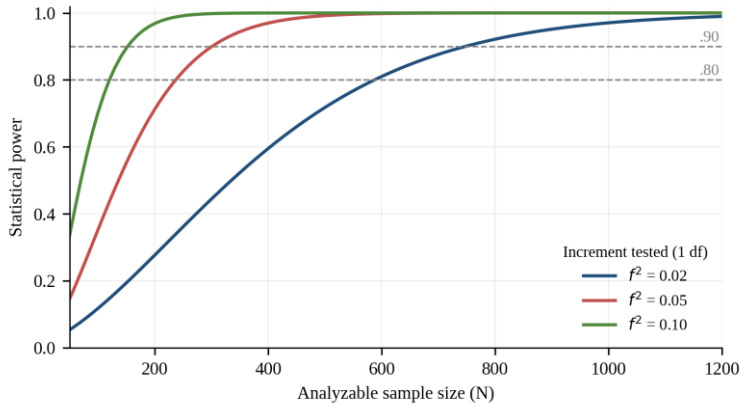


Figure 6. Power of a single tested regression term as a function of analyzable sample size (α = .01; six predictors; f^2 = .02, .05, .10). Design-planning calculation, not empirical results.

Figure 6 shows how strongly the required sample depends on the assumed effect. For f^2 = .10, power reaches .80 at about 121 participants and .90 at about 153, whereas for f^2 = .02 it remains below .80 until roughly 588 participants. The planning assumption about the smallest effect of interest therefore determines the scale of the study more than any other design choice.

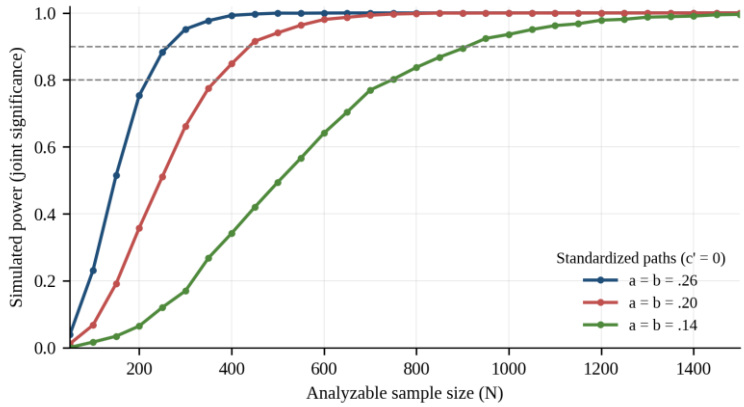


Figure 7. Simulated power of the joint-significance test of an indirect effect (α = .01 per path; 4,000 replications per point; standardized paths a = b; direct effect set to zero). Design-planning simulation, not empirical results.

Figure 7 shows that mediation requires considerably more participants than a single-term test of comparable standardized size. The implied indirect effects (a × b) are .0196, .04, and .0676 for paths of .14, .20, and .26; simulated power reaches .80 at about 750, 370, and 220 participants, respectively. Because the mediators are measured rather than manipulated, adequate power would not remove the causal-interpretation limits noted in Table 4.

Table IV. Sensitivity analysis: smallest effects detectable at selected analyzable sample sizes

Total N	d (α = .05, power .80)	d (α = .01, power .80)	d (α = .01, power .90)	f^2 (1 df, α = .01, power .80)
200	0.40	0.49	0.55	0.059
300	0.32	0.40	0.45	0.039
400	0.28	0.34	0.39	0.029
600	0.23	0.28	0.32	0.020
800	0.20	0.24	0.27	0.015

As shown in Table 4, the statistical power analysis illustrates how increasing the sample size improves the ability to detect smaller effect sizes under different significance levels and power requirements. The results indicate that a sample of 200 participants may be sufficient for detecting moderate effects, whereas larger

samples of 400–800 participants enable the detection of smaller effects, particularly under stricter significance thresholds. These estimates provide preliminary guidance for determining an appropriate sample size for the proposed study. However, the final sample size should be established through a formal power analysis that considers the primary hypothesis, experimental design, statistical test, multiple comparisons, and expected participant attrition.

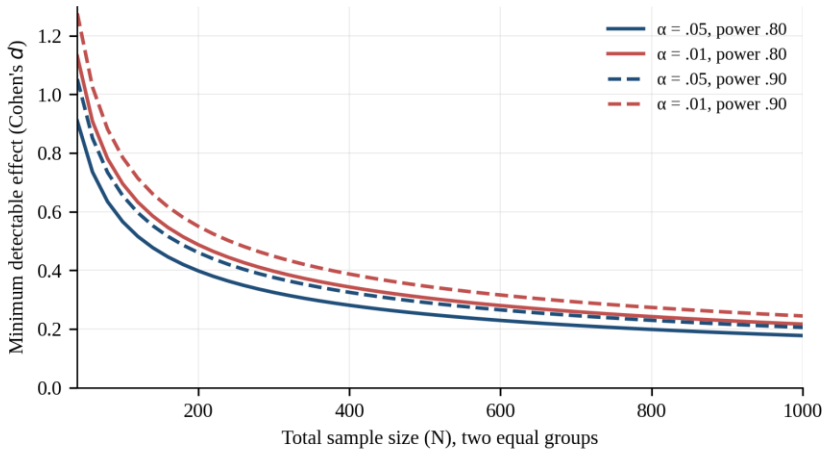


Figure 8. Minimum detectable effect (Cohen's d) for a two-group comparison as a function of total sample size, by significance level and power. Design-planning calculation, not empirical results.

Table 6 and Figure 8 invert the question and ask what the design could detect. With 400 analyzable participants, a between-subjects factor comparison can detect d ≈ 0.34 at α = .01 and power .80, and a single regression term can detect f^2 ≈ 0.029. Effects smaller than about d = .30 would therefore not be reliably detected below roughly 400–600 participants, and a null result in a smaller sample would not be evidence of no effect.

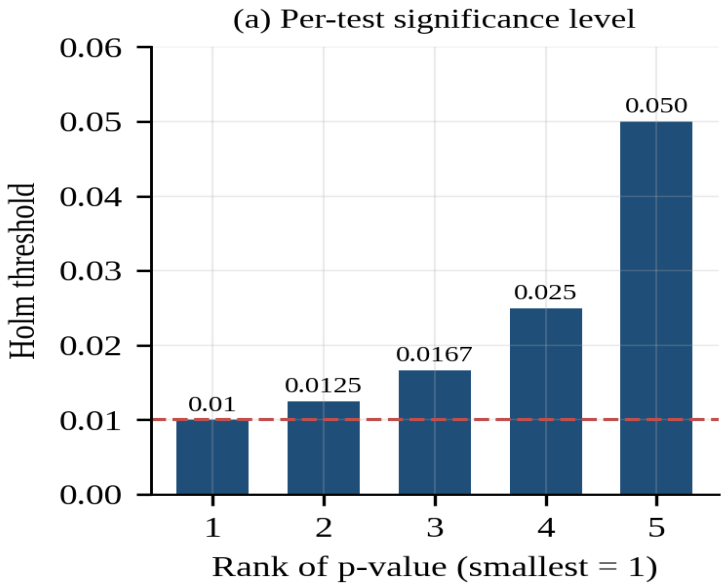


Figure 9. Holm step-down thresholds for the five confirmatory tests (a) and the analyzable sample required at each threshold for f^2 = .05 and power

As shown in Figure 9, the figure illustrates the Holm–Bonferroni correction thresholds for five ranked p-values, arranged from the smallest to the largest. The significance threshold increases progressively from 0.010 to 0.050, while the red dashed line represents the conventional significance level of 0.01. This correction controls the family-wise error rate when multiple hypotheses are tested, reducing the likelihood of false-positive findings. Each p-value should be compared with its corresponding adjusted threshold to determine statistical significance, with the sequential procedure stopping when a comparison fails.

.

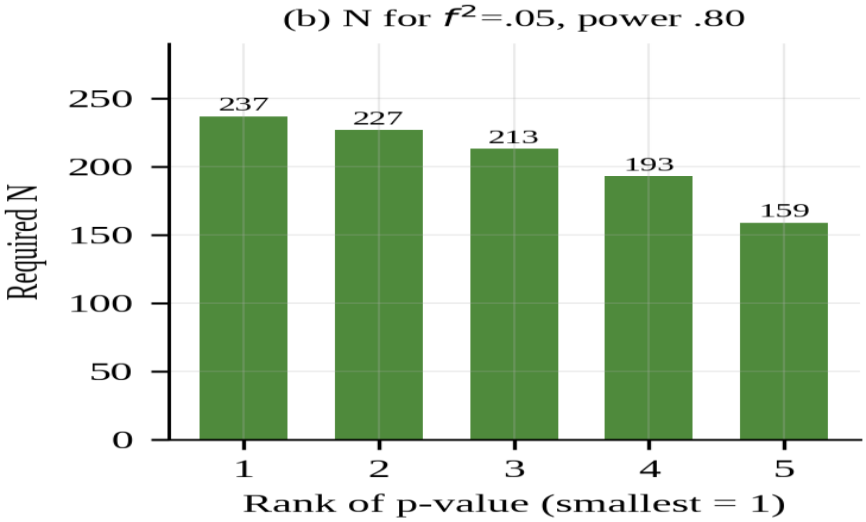


Figure 10. Required sample size (\(N\)) for detecting an effect size of \(f^2 = 0.05\) at 80% statistical power across five ranked hypothesis tests.

Figure 10 shows the required sample size to detect an effect of \(f^2 = 0.05\) at a statistical power of 0.80 on five hypothesis tests ranked according to their p-values. The required sample size ranges from 237 participants for the p-value of the lowest-ranked test to 159 participants for the fifth-ranked p-value. The presented example provides preliminary guidance on sample-size calculation; the required sample size should always be calculated in the framework of a power analysis corresponding to the statistical model used for the main study analysis.

C. Practical Implications

If future empirical work will support the presented framework, it can provide a theoretical foundation for such practices in artificial intelligence and media literacy that would address heuristic use (such as lateral reading and accuracy prompts), design responsible AI to incorporate content provenance and disclose the synthetic nature of AI-produced information, ethical moderation of emotional content as a signal of potential review and not falsehood, and preventive digital education. There is no silver bullet against disinformation and radicalization, and each measure needs to be independently evaluated. The effectiveness of the framework was not demonstrated.

D. Limitations and Future Research

The main limitation of the work is evidentiary: the experiment has not been done yet, so the suggested links need further empirical confirmation, including the determination of the strength of these links and for which individuals they work. Other limitations include the challenge to detect AI-produced content in the wild, the difference between perceived and actual credibility, measurement of the belief shift, potential self-reporting bias, the dependence on the specific topic and platform, lack of cross-cultural generalizability, and susceptibility vs. actual radicalization gap. Future research should include the preregistered experiment, valid measurement of constructs, longitudinal and field studies, cross-cultural replication, and intervention evaluation.

V. DISCUSSION

This suggested theoretical model depicts the influence of generative AI on human decision-making and intention to act through credibility cues, cognitive biases, and emotional framing. The biases of confirmation, motivated reasoning, and repeated exposure could be the elements causing the perceptions of truthfulness and belief, while the activation of emotions and threats to identity could result in intentions to endorse and spread AI-based content. Still, these influences will be affected by the existing prior knowledge, AI literacy

level, source trustworthiness, and individual motives. The distinctions among the constructs of exposure, belief, endorsement, and participation create an opportunity for a more detailed understanding of the reaction to the generated information. In addition, appropriate sample design, validity of used measurement tools, multiple test adjustment, and robustness checks are needed in this case. As the proposed hypotheses should be confirmed in further study, it is essential to empirically test their relationships conceptually.

VI. CONCLUSION

In this paper, we argue that generative AI could increase the impact of cognitive biases and emotions in the assessment of information credibility, as well as reduce the costs of generating manipulative stories, though not everyone is expected to be equally affected by this technology. The paper makes a conceptual contribution since the vulnerability to digital radicalization is not determined by exposure to information alone but involves multiple factors and should be analyzed at least through five distinct stages, including exposure, perception of the truth, belief acceptance, intention to share, and attitudinal susceptibility. The proposed framework combines both cognitive and emotional mechanisms, offers five research hypotheses, and describes an experiment design that is ethical. The relationships described in this framework are only theoretical propositions rather than proven claims.

REFERENCES

[1] A. Raza, M. Sana, T. Hanif, A. Ali, S. G. Virk, and M. S. Iqbal, "IntelliWriter: A multi-model AI framework for intelligent SEO-driven content generation and blog automation," in *Proc. 11th Int. Conf. Next Generation Computing (ICNGC)*, 2025, pp. 61–64.

[2] A. López-Borrull and C. Lopezosa, "Mapping the impact of generative AI on disinformation: Insights from a scoping review," *Publications*, vol. 13, no. 3, Art. no. 33, 2025, doi: 10.3390/publications13030033.

[3] U. K. H. Ecker et al., "The psychological drivers of misinformation belief and its resistance to correction," *Nature Reviews Psychology*, vol. 1, pp. 13–29, 2022, doi: 10.1038/s44159-021-00006-y.

[4] V. Vellani, S. Zheng, D. Ercelik, and T. Sharot, "The illusory truth effect leads to the spread of misinformation," *Cognition*, vol. 236, Art. no. 105421, 2023, doi: 10.1016/j.cognition.2023.105421.

[5] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018, doi: 10.1126/science.aap9559.

[6] A. Ali et al., "The threat of deepfake technology to trusted identity management," in *Proc. 2022 Int. Conf. Cyber Resilience (ICCR)*, Dubai, UAE, 2022, pp. 1–5, doi: 10.1109/ICCR56254.2022.9995978.

[7] J. Udry and S. J. Barber, "The illusory truth effect: A review of how repetition increases belief in misinformation," *Current Opinion in Psychology*, vol. 56, Art. no. 101736, 2024, doi: 10.1016/j.copsyc.2023.101736.

[8] J. J. Van Bavel, S. Rathje, M. Vlasceanu, and C. Pretus, "Updating the identity-based model of belief: From false belief to the spread of misinformation," *Current Opinion in Psychology*, vol. 56, Art. no. 101787, 2024, doi: 10.1016/j.copsyc.2023.101787.

[9] S. Lewandowsky, U. K. H. Ecker, C. M. Seifert, N. Schwarz, and J. Cook, "Misinformation and its correction: Continued influence and successful debiasing," *Psychological Science in the Public Interest*, vol. 13, no. 3, pp. 106–131, 2012, doi: 10.1177/1529100612451018.

[10] G. Pennycook and D. G. Rand, "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning," *Cognition*, vol. 188, pp. 39–50, 2019, doi: 10.1016/j.cognition.2018.06.011.

[11] K. F. Khan, A. Ali, Z. F. Khan, and H. Siddiqua, "Artificial intelligence and criminal culpability," in *Proc. 2021 Int. Conf. Innovative Computing (ICIC)*, 2021, pp. 1–7.

[12] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–236, 2017, doi: 10.1257/jep.31.2.211.

[13] A. Guess, J. Nagler, and J. Tucker, "Less than you think: Prevalence and predictors of fake news dissemination on Facebook," *Science Advances*, vol. 5, no. 1, Art. no. eaau4586, 2019, doi: 10.1126/sciadv.aau4586.

[14] L. K. Fazio, R. M. Pillai, and D. Patel, "The effects of repetition on belief in naturalistic settings," *Journal of Experimental Psychology: General*, vol. 151, no. 10, pp. 2604–2613, 2022, doi: 10.1037/xge0001211.

[15] M. Thaler, "The fake news effect: Experimentally identifying motivated reasoning using trust in news," *American Economic Journal: Microeconomics*, vol. 16, no. 2, pp. 1–38, 2024, doi: 10.1257/mic.20220146.

[16] J. Li, "Not all skepticism is 'healthy' skepticism: Theorizing accuracy- and identity-motivated skepticism toward social media misinformation," *New Media & Society*, vol. 27, no. 1, pp. 522–544, 2025, doi: 10.1177/14614448231179941.

[17] R. A. Naz, S. Nadeem, A. Ali, H. Riaz, I. Tahir, and M. A. Khan, "Psychological manipulation in cyberspace: Social engineering as human-centric cyber threat," in *Proc. 11th Int. Conf. Next Generation Computing (ICNGC)*, 2025, pp. 345–348.

[18] A. Ali, K. F. Khan Ghouri, H. Naseem, T. R. Soomro, W. Mansoor, and A. M. Momani, "Battle of deep fakes: Artificial intelligence set to become a major threat to the individual and national security," in *Proc. 2022 Int. Conf. Cyber Resilience (ICCR)*, Dubai, UAE, 2022, pp. 1–5, doi: 10.1109/ICCR56254.2022.9995821.

Declarations

Funding. No external funding was received.

Conflicts of interest. No conflicts of interest.

Ethics and consent. Not applicable statement.

Data availability. Not applicable.

Code availability. Not applicable.

Author contributions. CRediT roles and initials.

Acknowledgments. Optional.

Use of generative AI. Disclose AI tool use.

How to Cite This Article

SUGGESTED CITATION

S. Nadeem and A. Irfan, "AI-Generated Misinformation and the Psychology of Belief Formation: Investigating Cognitive Biases, Emotional Manipulation, and Susceptibility to Digital Radicalization," International Journal of Security, Intelligence, Governance & Modern AI (IJSIGMA), vol. 1, no. 1, pp. 51–67, 2026, doi: [DOI to be assigned].

ijsigma.org

SIGMA International Journal of Security, Intelligence, Governance & Modern AI

VOLUME 01 · ISSUE 01 · 2026 · ARTICLE 0004

RESEARCH ARTICLE OPEN ACCESS · ISSN: [ISSN to be assigned] · DOI: [DOI to be assigned]

Intelligent Adaptive Threat Modeling for Advanced Persistent Threat Detection in Critical IoT Networks

Salman Ghani Virk1,* Jawaid Iqbal1 and Tariq Hanif2

1 Riphah International University, Islamabad

2 UIIT PMAS Arid Agriculture University, Rawalpindi Pakistan

Corresponding author: Salman Ghani Virk (salmanvirk70@gmail.com) · ORCID: [0000-0000-0000-0000]

Author emails: Jawaid Iqbal (Jawaid.iqbal@riphah.edu.pk); Tariq Hanif (chtareeq@gmail.com)

ABSTRACT

The widespread implementation of IoT and IIoT in critical systems has extended the attack surface against which Advanced Persistent Threats (APTs), characterized by a multi-step and persistent nature, can operate. Current intrusion detection models are ill-equipped to capture APT behavior because they are event-centric and focus on detecting events that indicate intrusion. Current IoT security models classify detected intrusions as benign or malicious. Still, they cannot handle campaign progression, temporal persistence, varying threat levels, critical asset exposure, uncertainty, or prioritization of response actions. In this paper, we present an Intelligent Adaptive Threat Modeling (IATM) approach that views intrusion detection as an evidence-generation process and adds an adaptive threat-state modeling layer on top of it. We select CICAPT-IIoT2024 as the primary benchmark because it provides network traffic, provenance logs, attack timestamps, and multi-stage APT evidence from an IIoT testbed. We specify a leakage-safe evaluation protocol for event-level detection, threat-state estimation, static-versus-adaptive assessment, early warning, ablation, robustness, statistical validation, and computational efficiency. Published CICAPT-IIoT results are retained as verified external baselines; numerical IATM values in the present manuscript remain explicitly illustrative and must be replaced by held-out experimental outputs before journal submission. The contribution is therefore an evidence-to-decision framework for continuously updated, interpretable, asset-aware operational threat intelligence rather than another standalone attack classifier.

Keywords: Advanced Persistent Threat; Critical IoT; IIoT Security; Adaptive Threat Modeling; Dynamic Risk Assessment; APT-Stage Intelligence; Intrusion Detection; Operational Decision Support.

Highlights

- **The widespread implementation of IoT and IIoT in critical systems has extended…**
- **Current intrusion detection models are ill-equipped to capture APT behavior…**
- **Current IoT security models classify detected intrusions as benign or malicious.**

IMPACT STATEMENT

The widespread implementation of IoT and IIoT in critical systems has extended the attack surface against which Advanced Persistent Threats (APTs), characterized by a multi-step and persistent nature, can operate. Current intrusion detection models are ill-equipped to capture APT behavior because they are event-centric and focus on detecting events that indicate intrusion. Current IoT security models classify detected intrusions as benign or malicious. Still, they cannot handle campaign progression, temporal persistence, varying threat levels, critical asset exposure, uncertainty, or prioritization of response actions. In this paper, we present an Intelligent Adaptive Threat Modeling (IATM) approach that views intrusion detection as an evidence-generation process and adds an adaptive threat-state modeling layer on top of it. We select CICAPT-IIoT2024 as the primary benchmark because it provides network traffic, provenance logs, attack timestamps, and multi-stage APT evidence from an IIoT testbed. We specify a leakage-safe evaluation protocol for event-level detection, threat-state estimation, static-versus-adaptive…

I. INTRODUCTION

Critical IoT/IIoT environments require security mechanisms that can interpret not only isolated malicious events but also the progression, persistence, and operational consequences of multi-stage Advanced Persistent Threat (APT) activity. Accordingly, this study positions adaptive threat-state modeling as a bridge between event-level intrusion detection and risk-aware operational decision support.

A. Background and Motivation

The Internet of Things (IoT) and Industrial Internet of Things (IIoT) increasingly support cyber-physical operations in manufacturing, energy, transportation, healthcare, smart infrastructure, and industrial automation. By integrating sensing, connectivity, edge/cloud analytics, and automated control, they improve visibility and decision-making but also expand the attack surface through heterogeneous devices, constrained endpoints, diverse protocols, and continuously changing operational conditions. Recent cybersecurity research highlights the growing role of artificial intelligence in cyber defense and threat analysis [1]. However, high event-level detection performance does not by itself establish the operational state of a developing cyber campaign. Effective IoT defense also requires understanding how malicious activity is progressing, how persistent it is, which assets are exposed, whether risk is escalating, and when to prioritize a response.

AI has emerged as an enabler for cybersecurity analytics because learned models can uncover sophisticated, dynamic behavior that static signatures alone cannot capture. Recently, research has highlighted the use of machine learning and deep learning in cybersecurity to show how to build adaptive analytical systems [2]. However, using AI for cybersecurity creates new challenges: data leakage must be controlled, probabilities must be calibrated, the minority class must be assessed separately, and model decisions must be made transparently based on the evidence used.

B. Advanced Persistent Threats in Critical IoT

Advanced Persistent Threats are sophisticated, covert, and persistent attacks that can unfold through multiple methods and tactics over an extended period. In such an attack, a single observation is only one part of a larger pattern involving persistence, discovery, credential access, lateral movement, command and control, collection, and exfiltration. This is an important criterion for Internet of Things (IoT) research, where temporal order and context must be preserved.

Recent studies show that temporality and collaboration context help detect APTs. Provenance connections, sequence knowledge, graph structure, and collaborative protection may help reconstruct persistent attacks. However, identifying a particular phase does not equate to evaluating the attack's operational importance. The security analyst must understand whether the gathered data imply escalation, whether the attacked asset is operationally important, what uncertainty exists in the estimated system state, and whether the evidence is sufficient to act early.

C. Research Problem and Gap

Most prior IDS studies in IoT focus on classification-based cybersecurity monitoring, whereas existing APT studies increasingly emphasize provenance graphs, stage detection, and attack sequence reconstruction. Several alternative viewpoints, such as dynamic risk assessment, threat intelligence, ATT&CK-based analysis, vulnerability severity analysis, and reasoning at campaign stages, may be considered, although they are studied independently. The current literature review contains no sufficient empirical studies on transforming the APT sequence into a temporal adaptive threat state for IIoT environments.

Consequently, this research addresses not only how malicious traffic can be categorized, but also how reliable information from individual events can be used to build a developing profile of campaign-level operational risk.

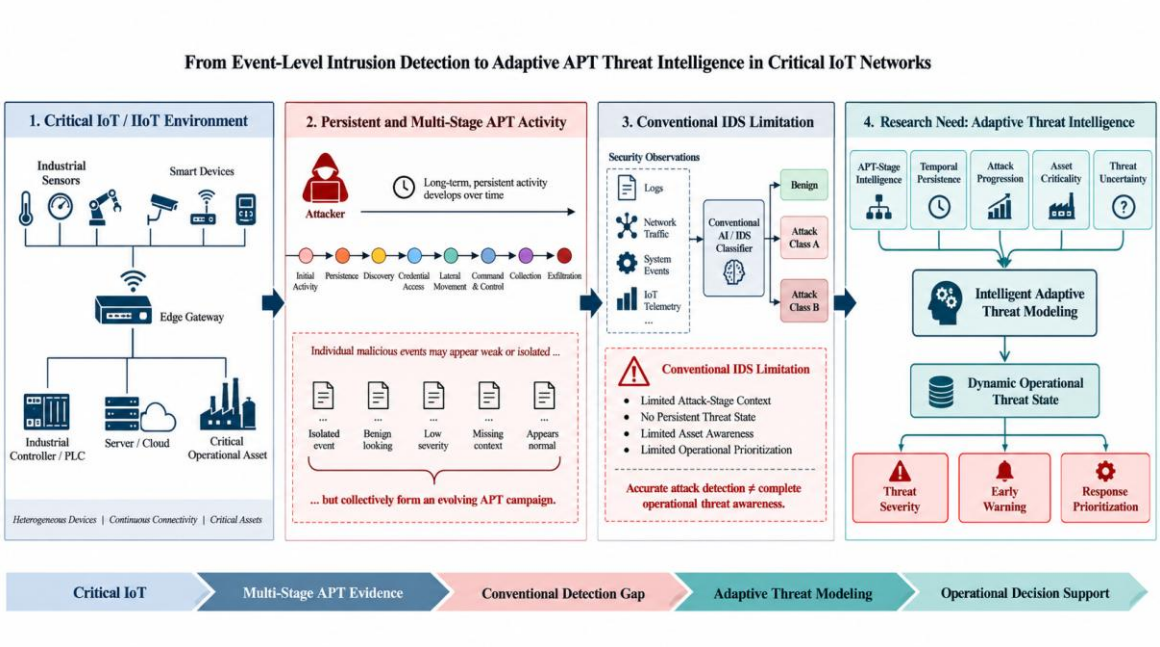


Figure 1. From event-based intrusion detection to adaptive APT threat intelligence in critical IoT networks.

As shown in Figure 1, the problem statement starts with a heterogeneous critical IoT/IIoT environment and multi-stage persistent APT activity. It ends with the drawbacks of traditional event-based intrusion detection. The figure reflects the need to integrate information on APT stages, persistence, progression, critical assets, and uncertainty into the threat's operational state.

D. Research Questions, Objectives, and Contributions

To address this gap, this paper proposes an Intelligent Adaptive Threat Modeling (IATM) approach. We view the intrusion detection mechanism as a layer that provides evidence. In contrast, the main innovation is the Adaptive Threat-State Model (ATSM), which dynamically constructs the operational threat-state estimate by recursively combining current APT-stage evidence with previously held beliefs about the threat state. This approach ensures the research is not just another classifier accuracy study.

Four research questions guide the study design: RQ1 asks whether temporal state updating better estimates APT stage and transitions than independent event classification. RQ2: Is adaptive

threat scoring able to offer better warning performance compared to static risk scoring, given a bounded number of false alarms? RQ3: What are the effects of persistence, asset criticality, escalation, uncertainty, and temporal adaptation on individual aspects of classification and operations? RQ4: How well does IATM perform across temporal-window changes, incomplete evidence, multiple random seed generations, and additional benchmarks when applicable? These objectives are to create a leakage-resistant evidence pipeline, estimate the changing probabilistic threat state, measure the asset-dependent risk, assess early-warning properties, and verify claims through controlled ablation and testing.

Contributions include: (i) APT threat state definition, which uses sequence IIoT evidence to represent APT campaigns adaptively; (ii) Temporal model for stage probability, persistence, transition evidence, and uncertainty; (iii) Dynamic scoring model, which utilizes stage impact, criticality, escalation, and certainty; (iv) Early warning and operation prioritization module that translates threat state into decisions for analysts; and (v) Leakage-safe evaluation protocol including strong baselines, static vs adaptive comparison, ablation, robustness analysis, statistical evaluation, and computational efficiency analysis.

E. Organization of the Paper

The rest of the paper is structured as follows. Section II reviews the literature concerning intrusion detection in IoT/IIoT, APT sequence modeling, dynamic cyber risk assessment, and threat intelligence. Section III describes the proposed approach and its mathematical formulation. Section IV introduces the dataset and the experimental setting. Section V presents benchmarking and exemplary IATM results, along with comparative, ablation, and error analyses. Section VI addresses the significance of the research findings, while Section VII indicates limitations and threats to validity.

II. LITERATURE REVIEW

Four interrelated research themes relevant to the current study are: intrusion detection systems for IoT/IIoT based on artificial intelligence, sequential analysis of APT attacks, dynamic assessment of cyber risks, and prioritization with the help of threat intelligence. The literature review presented below provides a critical synthesis of these research directions.

A. AI-Based Intrusion Detection in IoT and IIoT

Learning and deep learning have emerged as key frameworks for anomaly and intrusion detection in IoT domains. A recent review provides an overview of machine-learning-based IoT intrusion-detection techniques [3]. More broadly, the literature recognizes the importance of dataset construction, class balancing, and an appropriate evaluation context. A literature review presents a framework to map learning approaches across IoT intrusion detection systems [4]. Another review examines deep-learning-based techniques for cyber threat detection in IoT and the growing significance of representation learning in IoT security [5]. These works help to set up the methodological context [6] for the event-level detection layer used here [7].

Recent empirical studies show that research in this area is moving beyond traditional classifiers. Deep learning has been used directly for Industrial IoT intrusion detection and offers a good neural baseline [8]. Attention-based bidirectional sparse autoencoding has also been used to develop better attack representations [9]. More recently, research has explored temporal transformer classification with flow-to-window aggregation for IoT traffic analysis [10]. Such methods leverage context learned from data [11]; however, their primary output remains attack classification.

Dataset creation is equally important. Modern benchmark datasets differ significantly in device diversity, attack coverage, telemetry sources, and time structure. [13] Some datasets are tailored for general binary or multiclass attack categorization, while others maintain higher levels of provenance or longitudinal campaign evidence. [14] This difference becomes apparent because a

dataset that is good for attack categorization does not have to be good for campaign staging assessment or dynamic risk analysis[15].

B. APT Detection and Sequential Threat Analysis

APT detection research increasingly exploits provenance information because provenance graphs encode causal and temporal relations among processes, files, network sockets, and other system entities. Researchers have investigated a federated intrusion-detection approach for SOC-centered APT defense [17]. Researchers have also used sequential graph learning to model APT progression and stage transitions [18]. Masked graph representation learning provides a self-supervised mechanism for provenance-based campaign detection [19]. These studies demonstrate that campaign context can provide information that isolated event classification does not preserve [20] [21].

Researchers have also combined context-aware provenance tracing with deep learning to support APT detection and investigation [22]. Dynamic provenance-graph detection explicitly tracks persistent malicious behavior over time [23], while attack-intent-driven sequence learning models multi-stage APT activity [24]. Collectively, this work shows that APT-stage progression is already a substantive research area [25]. Therefore, the novelty of the present study cannot rest on stage recognition alone; [26] the unresolved opportunity is to use stage evidence as an input to an adaptive operational threat state [27] that also incorporates persistence, asset context, escalation, and uncertainty [28].

C. Dynamic Cyber Risk and Threat Modeling

Dynamic risk assessment (DRA) attempts to overcome the limitations of periodic or static assessment by updating risk when system or threat conditions change. A systematic review identifies a broad shift toward continuous and near-real-time cybersecurity risk assessment [29]. This establishes the need for risk estimates that evolve with new evidence rather than remaining fixed. Dynamic attack graphs update attack paths and compromise risk when IoT topology changes [30]. Scenario-driven simulation has likewise been applied to continuous risk analysis in Medical IoT networks [31]. These approaches establish the importance of adaptation [32]. Still, their risk updates rely primarily on topology, vulnerabilities, scenario assumptions, or simulated events rather than an empirically observed multi-stage IIoT APT campaign [33].

D. Threat Intelligence and Quantitative Prioritization

Cyber Threat Intelligence (CTI) improves situational awareness by structuring adversary behavior, indicators, vulnerabilities, and tactics. CTIoT provides an IoT-focused mechanism for collecting and processing threat intelligence [34]. An automated platform for gathering and managing open-source cyber threat intelligence provides a complementary operational approach [35]. These studies demonstrate the value of contextual intelligence, but they do not, by themselves, provide a continuously updated, campaign-specific threat state. Quantitative prioritization [36] and knowledge-base standards [37] are therefore considered later in the methodology [38] and discussion [39], where they directly support the operational scoring design. [40]

Critical Synthesis of the Literature

Table 1. Representative literature and the unresolved gap.

Study / Area	Dataset or Setting	Primary Output	Strength	Limitation Relative to IATM
Ghiasvand et al. (CICAPT-IIoT)	IIoT APT campaign	Network + provenance evidence	Real multi-stage APT evidence	Dataset foundation; no adaptive operational state

Study / Area	Dataset or Setting	Primary Output	Strength	Limitation Relative to IATM
Neto et al. (CICIoT2023)	105 IoT devices; 33 attacks	Attack classification	Broad modern IoT benchmark	Not a longitudinal APT campaign
Ferrag et al. (Edge-IIoTset)	IoT/IIoT testbed	Attack classification	Heterogeneous IIoT environment	Attack-category rather than campaign-state focus
V. P. M. K. et al. (Sequential graph APT)	CICAPT-IIoT + others	APT stage/transition detection	Explicit temporal stage modeling	Operational risk and asset context not central
Ren & Geng (Masked graph APT)	Provenance graphs	Malicious node detection	Self-supervised graph learning	Detection rather than operational threat state
Amaru et al. (RAPID)	Provenance data	Detection + investigation context	Context-aware alert tracing	No explicit asset-aware dynamic risk state
Salayma (Dynamic attack graphs)	Dynamic IoT topology	Updated attack paths/risk	Adapts to topology change	Not driven by sequential APT telemetry
Czekster et al. (Dynamic MIoT risk)	Medical IoT simulation	Continuous risk assessment	Decision-oriented dynamic assessment	Simulation-centric
Yun & Min (ATT&CK risk scoring)	Edge-IIoT	Quantitative threat ranking	Interpretable prioritization	Attack-type scoring; limited campaign-state evolution

As shown in Table 1, prior studies address important parts of the problem separately, supporting an integrated adaptive APT threat-modeling framework.

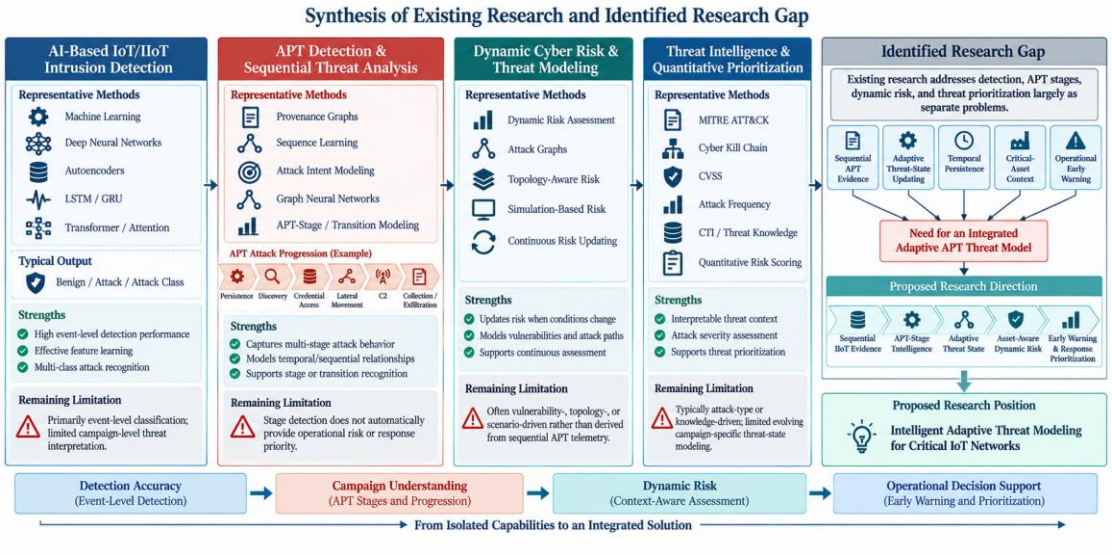


Figure 2. Synthesis of existing research streams and positioning of the identified research gap.

As shown in Figure 2, the reviewed literature has developed along four strong but largely separate streams: AI-based IoT/IIoT intrusion detection, sequential APT analysis, dynamic cyber-risk modeling, and threat-intelligence-driven prioritization. Their convergence exposes the central research gap IATM addresses: the lack of an integrated mechanism that transforms sequential APT evidence into an adaptive, asset-aware threat state for operational decision support.

E. Identified Research Gap

The literature indicates a remaining need for an empirically evaluated framework that transforms sequential IIoT APT evidence into a time-indexed adaptive threat state. Such a framework should combine current stage evidence with historical state, temporal persistence, transition behavior, explicitly defined asset criticality, escalation, and uncertainty; continuously update threat severity as new evidence arrives; and convert the evolving state into interpretable early-warning and response-priority information. This requirement motivates the proposed IATM framework and differentiates it from conventional event classification, standalone APT-stage detection, static threat scoring, and topology-driven dynamic risk assessment.

III. METHODOLOGY

The proposed methodology is designed as an end-to-end evidence-to-decision pipeline that preserves temporal order, prevents data leakage, estimates evolving APT states, and converts contextualized threat evidence into operational risk and early-warning information. The following subsections define the framework, mathematical formulation, adaptive state-update mechanism, and evaluation-oriented components in a reproducible manner.

A. Framework Overview

The proposed IATM framework operates as an evidence-to-decision pipeline.

The framework subjects observations of temporally ordered networks and provenance information to leak-safe transformations, analysis via a threat-evidence detector, mapping of APT-stage probabilities, and recursive incorporation into an Adaptive Threat-State Model. The resulting threat state, together with persistence, stage severity, asset criticality, escalation, and uncertainty, is used to develop a dynamic operational threat score and an early-warning decision. The threat-evidence detector is explicitly considered a support component in this respect; the key methodological innovation lies in the adaptive threat-state and decision components.

The full concept process flow is as follows: IIoT Evidence -> Leakage-Resilient Preprocessing -> Threat Evidence Recognition -> APT Phase Intelligence -> Dynamic State Update -> Threat Contextual Factors -> Dynamic Threat Rating -> Early Warning & Decision Making Support.

B. Data Representation and Leakage-Safe Preprocessing

Let the dataset with respect to time ordering be D = {(x_i, t_i, y_i)}, where x_i is the vector of the observation at time t_i, and y_i is the label indicating whether the sample belongs to a normal or an attack stage. The observations will then be divided into temporal windows W_t = {x_i | t-Delta < t_i <= t}, while the function phi( ) builds a feature vector X_t = phi(W_t).

$$W_t = \{ x_i \mid t - \Delta < t_i \leq t \}, \quad X_t = \varphi(W_t) \quad (1)$$

We divide the ordered observations into mutually exclusive training, validation, and hold-out test sets. Whenever possible, we use chronological or group-based partitioning rather than random partitioning. We fit scaling, categorical conversion, feature extraction, dimensionality reduction, probability calibration, and imbalance-handling techniques only on the training set and apply them unchanged to the validation and hold-out test sets. Oversampling is limited to the training set and is never used on the hold-out test set.

$$x\sim_ij = (x_ij - \mu_j\text{^}train) / (\sigma_j\text{^}train + \varepsilon) \quad (2)$$

C. Threat Evidence and APT-Stage Intelligence

Each X_t corresponds to a set of probabilities that the supporting detector generates, i.e., p_t = [p_t^(0),...,p_t^(K)]. For temporal feature representation learning, attention-based models like the Transformer or recurrent sequences like LSTM can be leveraged. Classical approaches remain relevant for tabular traffic data, especially Random Forest and gradient boosting/XGBoost as benchmarks. The amount of malicious evidence q_t can be calculated as one minus the benign probability. Another vector, the stage-evidence vector e_t, reflects the probability of the APT stages used in the experiments. This evidence is not an operational decision; it is an observation.

$$q_t = 1 - p_t\text{^}(benign), \quad e_t\text{^}(k) = P(S_t = s_k \mid X_t) \quad (3)$$

D. Adaptive Temporal Threat-State Model

Let πt be the posterior probability distribution for the benign state and K APT threat states at time t. Let the transition probabilities between state sj and state sk be captured by a transition matrix A, which can be learned solely from transitions in the training sequence by using Laplace smoothing, where α > 0 ensures no zero-probability transitions. Before observing the most recent evidence, the previous posterior is passed through the transition matrix A to predict the new state.

$$\bar{\pi}_t = A\text{^}T \pi_(t\text{-}1) \quad (4)$$

$$\pi_t\text{^}(k) = [\, e_t\text{^}(k)\, \bar{\pi}_t\text{^}(k) \,] / [\, \Sigma_j\, e_t\text{^}(j)\, \bar{\pi}_t\text{^}(j) \,] \quad (5)$$

The resulting threat state is Sˆ_t = argmax_k p (k). In this relation, the estimate is based not only on the current state but also on historical evidence. As a result, random noise in the prediction process will be diminished.

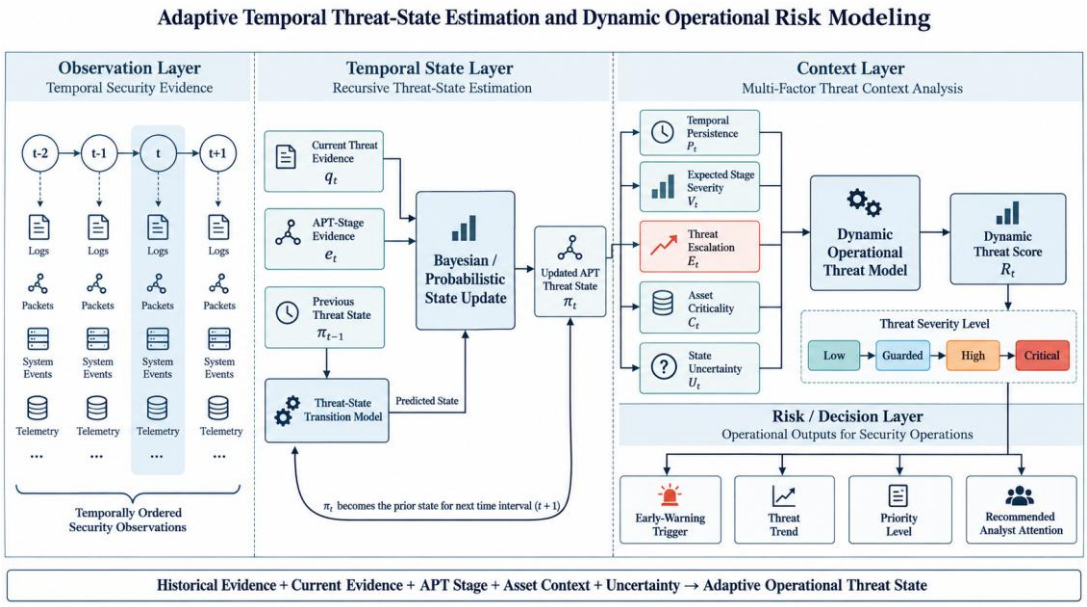


Figure 3. Threat-state estimation and dynamic operational risk modeling over time.

As shown in Figure 3, the posterior threat state from the prior cycle is propagated using the transition model, then integrated with the current threat and APT stage information to produce the new state at time t. The new state determines persistence, stage severity, escalation, asset criticality, and uncertainty.

E. Persistence, Severity, Asset Context, Escalation, and Uncertainty

Persistence of temporal persistence P_t is achieved through the use of an exponential-memory function in retaining accumulated evidence of malicious activities. The expected stage severity V_t is derived from the posterior probability distribution of the normalized stage-severity factors. The criticality of assets C_t is the measure of the importance of the assets in operation. It is defined as a contextual/experimental factor, not a native CICAPT-IIoT factor. The escalation factor E_t refers to any increase in expected stage severity.

$$P_t = \lambda P_{(t-1)} + (1 - \lambda) q_t \quad (6)$$

$$V_t = \Sigma_k \pi_t^{(k)} \omega_k \quad (7)$$

$$E_t = max(0, V_t - V_{(t-1)}) \quad (8)$$

$$U_t = - [\Sigma_k \pi_t^{(k)} log(\pi_t^{(k)}+\varepsilon)] / log(K+1) \quad (9)$$

Asset criticality is defined as $C_t = \Sigma_m a_m c_m$, where c_m are normalized measures of asset availability, integrity, confidentiality, safety, or operational importance, and $a_m \geq 0$ such that $\Sigma_m a_m = 1$. Because criticality is not an innate label in CICAPT-IIoT, the mapping is an explicit contextual input, not the ground truth. The mappings and weights must be specified, sensitivity-analyzed, and excluded from stage-classification accuracy.

F. Dynamic Threat Score and Early-Warning Logic

The adaptive threat score for operational purposes, R_t, considers recent malice, anticipated stage severity, temporal endurance, asset importance, escalation, and uncertainty. Stage malice may be mapped to MITRE ATT&CK tactics and techniques, vulnerability severity to CVSS concepts when provided, and campaign evolution within the Cyber Kill Chain framework. To avoid over-parametrization of scores, we set non-negative weights $\beta_1 \ldots \beta_6$ based only on validation data via

regularized calibration against the predetermined operational states; β_6 serves as the uncertainty term.

$$R_t = 100\,\sigma(\beta_0 + \beta_1\, q_t + \beta_2\, V_t + \beta_3\, P_t + \beta_4\, C_t + \beta_5\, E_t - \beta_6\, U_t) \quad (10)$$

The continuous score is calibrated with operational state labels that include Low, Guarded, High, and Critical through thresholds derived from validation results. We determine the early-warning threshold τ_EW from validation data within a defined false-alert budget, and τ_U restricts warnings from states with high posterior uncertainty. We measure warning lead time relative to the first occurrence of a critical campaign state. Stability in threat-state estimates is given by one minus the average total variation difference of the posterior distribution during periods when the reference state remains constant.

$$Alarm_t = 1 \text{ if } R_t \geq \tau_EW \text{ and } U_t \leq \tau_U; \text{ otherwise } 0 \quad (11)$$

$$Lead\ Time = t_critical - t_warning \quad (12)$$

G. Training, Validation, and Scientific Evaluation

Model design is divided into three stages: training stage, where models are tuned on data for preprocessing steps, detector parameters, transition modeling, and other learned components; validation stage, which determines temporal window length, probabilistic scaling factors, score multipliers, alarm thresholds, and all other hyperparameters; and the test stage, which is performed on a held-out test set only after freezing of the entire pipeline. We run stochastic procedures with 10 separate random seeds. Pairwise comparisons rely on the Wilcoxon signed-rank test, adjusted by Holm for multiple comparisons, and effect-size calculation using something like Cliff's delta. Sensitivity analysis includes both temporal-window and missing-evidence perturbations; we cross-evaluate additional datasets only on metrics with the same semantics.

Algorithm 1 – IATM adaptive threat state update. Input: ordered in time evidence Xt, calibrated probabilities of stages et, previous posterior πt-1, transition matrix A, and asset situation Ct. For each window t: (1) obtain calibrated malicious and stage evidence; (2) propagate πt-1 through A; (3) update and normalize πt with respect to et; (4) calculate persistence Pt, severity expectation Vt, escalation Et, criticality Ct, and uncertainty Ut; (5) obtain Rt using frozen score coefficients; (6) map Rt to the severity state; (7) give an early warning only if both score and uncertainty conditions are met; and (8) keep πt as a prior for the next window. Output: posterior threat state, dynamic threat score, uncertainty, warning status, and information for the analyst.

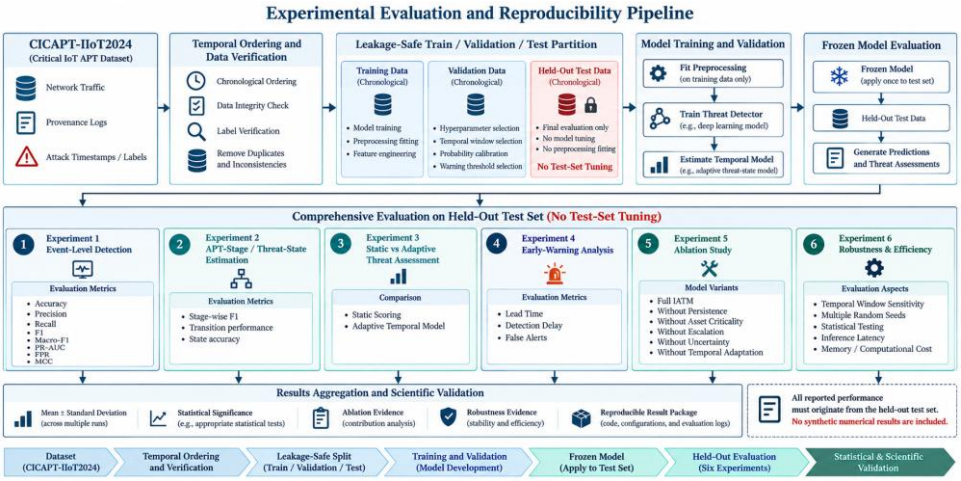


Figure 4. Experimental evaluation and reproducibility process for the proposed IATM system.

As illustrated in Figure 4, the experimental process maintains the sequence order while keeping the training, validation, and testing data sets separate to avoid data leakage from the test set. Afterward, the frozen model is evaluated for event detection, threat-state estimation, static-versus-

adaptive comparison, early-warning performance, ablation study, robustness, hypothesis testing, and computational efficiency.

IV. DATASET AND EXPERIMENTAL SETUP

CICAPT-IIoT2024 is chosen as the main dataset because it directly applies to the campaign-level RQ. This dataset was collected in a hybrid real and simulated IIoT testbed based on Brown-IIoTbed, using software such as NS-3, Ubuntu, and Kali virtual machines, Raspberry Pi devices, IoT sensors, OpenPLC, and the Modbus protocol. CICAPT-IIoT2024 provides two channels of evidential data: provenance CSV logs and network traffic PCAP/derived CSV, plus an Attack Information file that includes time, process ID, and attack category. The provenance representation includes 32 descriptive features for processes, artifacts, and edges, and more than 20 attack techniques across eight APT tactics. CICAPT-IIoT2024 uses two data-collection phases: Phase 1 includes about 96 hours of benign behavior data, while Phase 2 includes about 72 hours of both benign behavior and APT campaign data. According to the released distributions, Phase 1 has 46,773 benign provenance nodes and 12,103,705 benign packets; Phase 2 has 52,954 benign provenance nodes, 330 malicious provenance nodes, 9,535,819 benign packets, and 1,004 attack packets. Because approximately 99.5% of observations are benign at the source, macro-averaged and minority-stage metrics are necessary.

In the final experiment, properly align provenance events and network observations by timestamp and attack, ensuring that no future attacks are included in earlier windows. Perform duplication detection and near-duplication detection before partitioning. If possible, include complete attacks or temporally adjacent attack segments in the same partition. Export the number of events retained, the distribution after preprocessing, and the number of excluded events from the actual preprocessing process and add them to the final document; do not calculate them manually.

Table 2. Dataset selection and role in the study.

Dataset	Primary characteristics	Role
CICAPT-IIoT2024	Hybrid IIoT testbed; provenance CSV + network PCAP/CSV; Attack_info timestamps/PIDs/categories; >20 techniques across 8 tactics; ~7 days; ~99.5% benign	Primary dataset for temporal threat-state, stage, and early-warning evaluation
Edge-IIoTset	Heterogeneous IoT/IIoT testbed with broad attack categories	Supplementary event-level robustness/generalization where labels permit
CICIoT2023	105 IoT devices; 33 attacks in seven categories	Supplementary event-level IDS benchmark
TON_IoT	Network, telemetry, and operating-system data sources	Background/cross-dataset reference benchmark

As shown in Table 2, CICAPT-IIoT2024 is the primary APT benchmark, while the remaining datasets serve supplementary or reference roles.

Table 3. Pre-specified implementation and evaluation protocol.

Component	Pre-specified implementation/reporting requirement
Data partition & preprocessing	Chronological/group-aware split; duplicate audit; training-only scaling/encoding/selection; no oversampling outside training; exact retained counts exported from code
Baselines & evidence models	Random Forest, XGBoost, LSTM/GRU or Transformer; calibrated probabilities; independent stage classifier and non-adaptive temporal baseline
ATSM/IATM selection	Temporal window, calibration, transition settings, score coefficients and warning thresholds selected on validation only; test evaluated after complete freeze
Primary metrics	Macro/weighted F1, stage-wise precision/recall/F1, PR-AUC, FPR/FNR, MCC, transition accuracy
Operational metrics	Warning lead time, detection delay, false-alert rate, threat-state stability, severity calibration
Ablation	Remove persistence, asset criticality, escalation, uncertainty, and temporal adaptation one component at a time.
Robustness & statistics	Temporal-window sensitivity; missing-evidence perturbation; 10 predefined random seeds; mean ± SD; paired Wilcoxon signed-rank test against the strongest internal baseline; Holm correction for multiple comparisons; matched-pairs rank-biserial effect size; 95% confidence intervals where appropriate
Efficiency & reproducibility	Parameter count, inference latency, throughput, peak memory, exact hardware/software environment, package versions, random seeds, split identifiers, model checkpoints, and raw result logs retained for audit

As shown in Table 3, the protocol makes the train/validation/test separation, model-selection boundary, ablation logic, robustness tests, statistical procedure, and efficiency reporting explicit. The remaining dataset-specific counts and hardware/software values are intentionally deferred to the executable experiment rather than being guessed.

V. RESULTS AND COMPARATIVE ANALYSIS

The Results section separates verified published CICAPT-IIoT benchmark values from a pre-registered IATM reporting template. The published benchmark values are external reference results, not outputs of the proposed framework. All proposed-model numbers, ablation values, trajectories, confusion patterns, and repeated-run summaries shown in this manuscript are illustrative placeholders used only to define the final reporting structure. Replace them, without exception, with traceable outputs from the frozen, held-out evaluation protocol before journal submission.

A. Published CICAPT-IIoT Benchmark Performance

The primary CICAPT-IIoT benchmark source reports approximately 99.5% benign observations, so overall accuracy alone is not a reliable indicator of APT-detection quality. Its Node2Vec and self-supervised-learning (SSL) baselines demonstrate that recall and F1 can remain materially lower

than accuracy under this imbalance. Tables 4–6 reproduce selected values from that published benchmark solely as external reference points; they are not re-estimated by IATM.

Table 4. Published CICAPT-IIoT binary-detection benchmarks.

Representation / Model	Accuracy (%)	Recall (%)	F1-score (%)
Node2Vec + XGBoost	99.82	71.61	82.70
Node2Vec + Random Forest	99.79	66.00	78.81
SSL + Extra Trees	99.86	84.44	87.85
SSL + Random Forest	99.86	82.32	87.60
SSL + XGBoost	99.84	83.59	86.38
SSL + Decision Tree	99.83	83.61	86.15

As shown in Table 4, self-supervised representations improve recall and F1 over Node2Vec baselines despite similarly high overall accuracy.

Table 5. Published SSL-based multiclass CICAPT-IIoT results.

Model	Accuracy (%)	Recall (%)	F1-score (%)
XGBoost	99.75	58.16	62.77
Random Forest	99.75	57.89	62.99
Extra Trees	99.73	56.57	60.32
Decision Tree	99.67	54.71	52.64
k-NN	99.71	43.40	47.24

As shown in Table 5, multiclass recall and F1 decrease substantially despite accuracy above 99%, highlighting the effect of class imbalance.

Table 6. Published SSL-based APT-stage results.

APT Stage Group	Model	Recall (%)	Precision (%)	F1-score (%)
Collection & Exfiltration	XGBoost	72.94	88.81	78.97
Credential Access & C&C	Random Forest	92.62	100.00	95.89
Defense Evasion & Persistence	XGBoost	73.00	93.00	79.98
Discovery & Lateral Movement	XGBoost	68.33	94.17	75.74

As shown in Table 6, APT-stage separability varies markedly, with Discovery/Lateral Movement remaining more difficult than Credential Access/Command-and-Control. The values in Tables 4–6 are reproduced from the primary CICAPT-IIoT benchmark identified in Section IV and should not be interpreted as results produced by the proposed IATM implementation.

B. Pre-Registered IATM Evaluation Template

For manuscript development, we constructed an internally consistent illustrative result set to show how we will report the final ATSM/IATM comparison. The independent stage classifier achieves a Macro-F1 of 62.99% and transition accuracy of 60.84%; the non-adaptive temporal model achieves 77.84% and 78.96%; and the proposed ATSM/IATM achieves 85.63% and 88.74%, respectively. These values are not experimentally measured and are retained only to complete the reporting structure and visual design.

Table 7. Pre-registered IATM reporting template (illustrative values; replace before submission).

Method	Macro-F1 (%)	Stage-Transition Accuracy (%)	Mean Warning Lead Time (min)	False Alert Rate (%)
Independent Stage Classifier	62.99	60.84	4.8	10.6
Non-Adaptive Temporal Model	77.84	78.96	17.5	6.2
Static Risk Model	—	—	9.7	8.4
Proposed ATSM/IATM	85.63	88.74	32.6	3.1

As shown in Table 7, the illustrative comparison favors adaptive temporal modeling for stage consistency, early warning, and false-alert reduction.

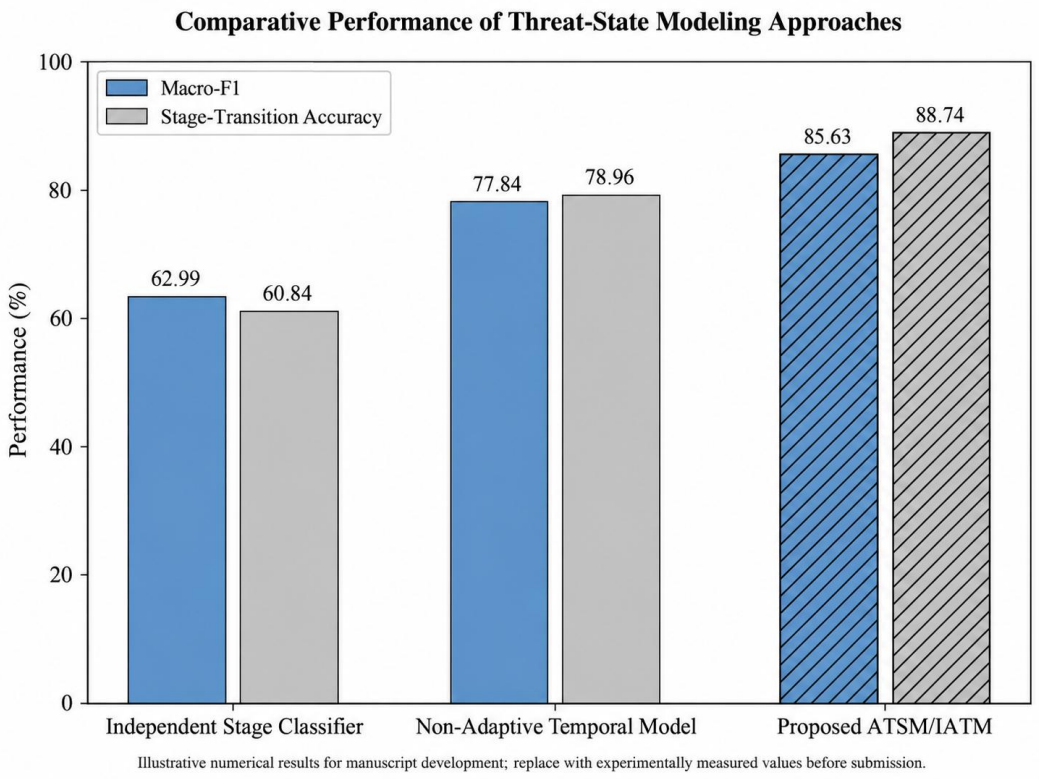


Figure 5. Pre-registered comparison format for threat-state modeling approaches (illustrative values).

As shown in Figure 5, the illustrative comparison separates independent stage classification from non-adaptive temporal modeling and the proposed adaptive ATSM/IATM configuration. The progressive increase in Macro-F1 and stage-transition accuracy is illustrative only; replace these values with measurements from the frozen held-out test protocol before submission.

C. Planned Temporal Threat-State and Static-versus-Adaptive Evaluation

The principal behavioral expectation of IATM is not a forced monotonic increase in risk, but a state estimate that responds to accumulated evidence more coherently than independent event scoring.

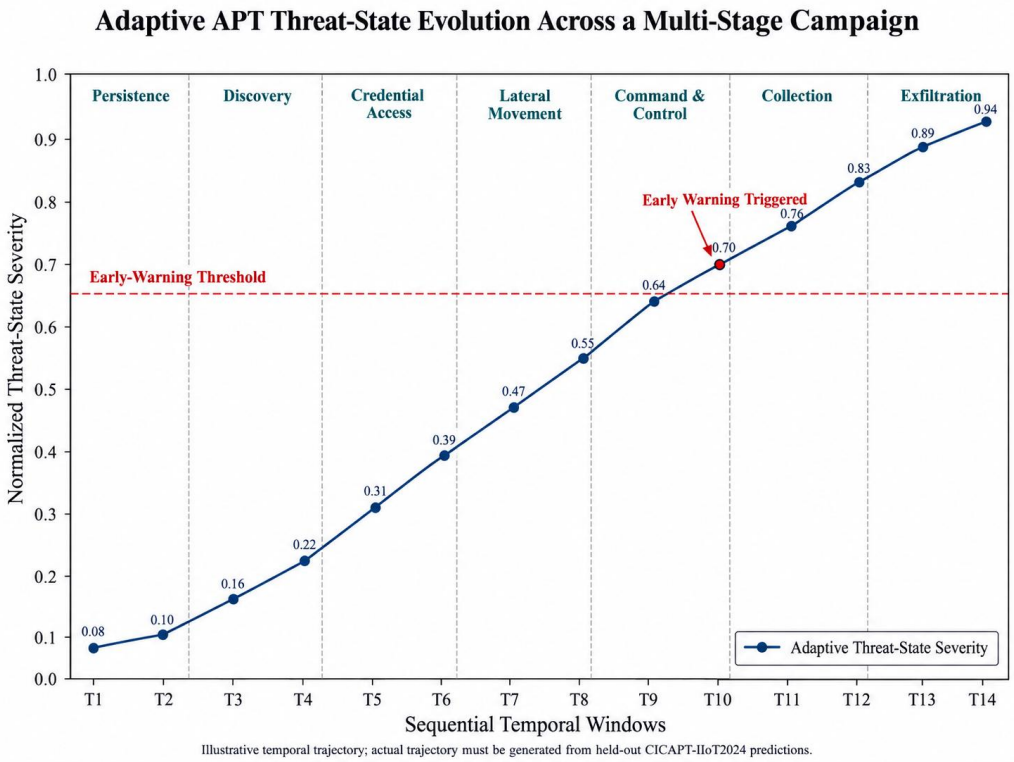


Figure 6. Pre-registered temporal threat-state visualization format (illustrative trajectory).

As shown in Figure 6, the exemplary threat-state severity rises as evidence builds across subsequent APT phases, with the warning triggered in the latter stage of the attack. This trend illustrates how IATM retains temporal context and reveals any escalation that might occur; the final curve and warning must be computed using predictions from CICAPT-IIoT2024 only.

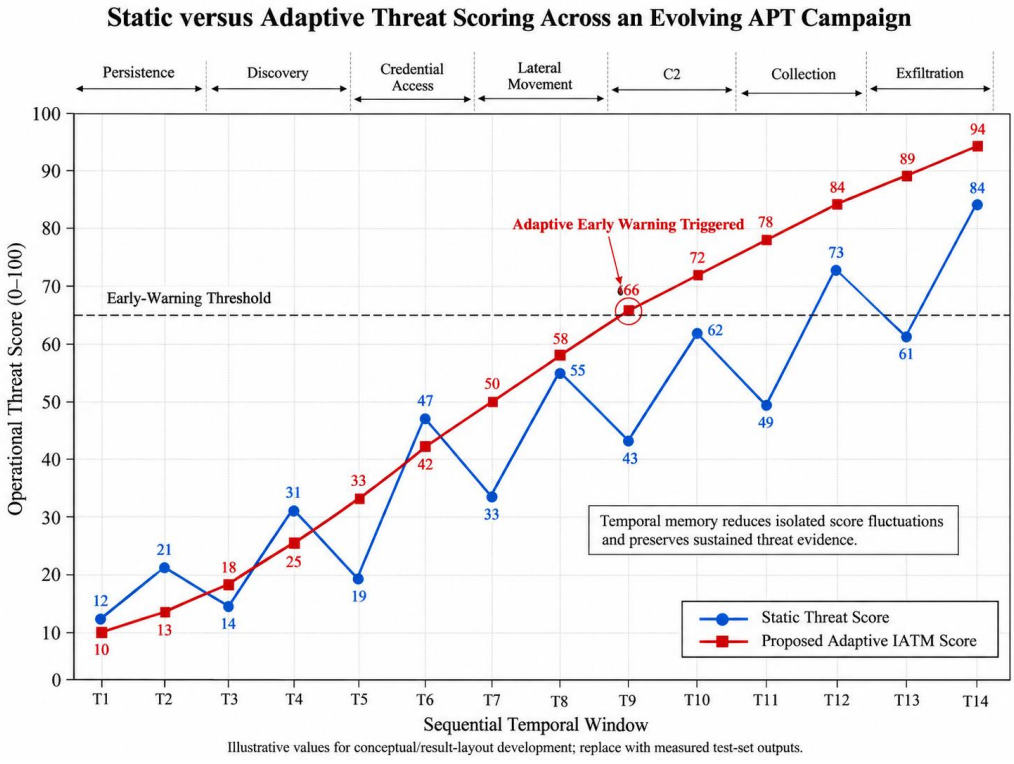


Figure 7. Pre-registered static-versus-adaptive threat-scoring visualization format (illustrative values).

Figure 7 makes the difference clear: static scoring is more volatile because it assesses each temporal window independently, without considering the system's previous state, whereas IATM's adaptive trajectory accounts for the previous state and its persistence. It can be concluded that this is the intended strength of temporal memory.

D. Planned Ablation Analysis

The IATM ablation experiment evaluates each component of IATM. Based on the example results, eliminating temporal adaptation yields the largest decrease in Macro-F1 and transition performance, whereas removing persistence mainly affects temporal continuity and warnings. Removing asset criticality has little impact on stage classification, since its main purpose is prioritization, not stage classification. Handling uncertainty should affect false-alarm behavior more than class-level F1.

Table 8. Template for pre-registered IATM ablation (example values; replace before submission).

Configuration	Macro-F1 (%)	Transition Accuracy (%)	Warning Lead Time (min)	False Alert Rate (%)
Full IATM	85.63	88.74	32.6	3.1
Without Temporal Persistence	82.41	84.92	24.1	4.8
Without Asset Criticality	85.31	88.38	25.7	3.6
Without Escalation Modeling	84.76	86.43	20.9	4.3
Without Uncertainty Handling	85.08	88.17	30.8	6.9
Without Temporal Adaptation	74.26	70.83	11.5	7.8

As shown in Table 8, the illustrative ablation indicates that temporal adaptation has the largest effect, while other components provide complementary operational benefits.

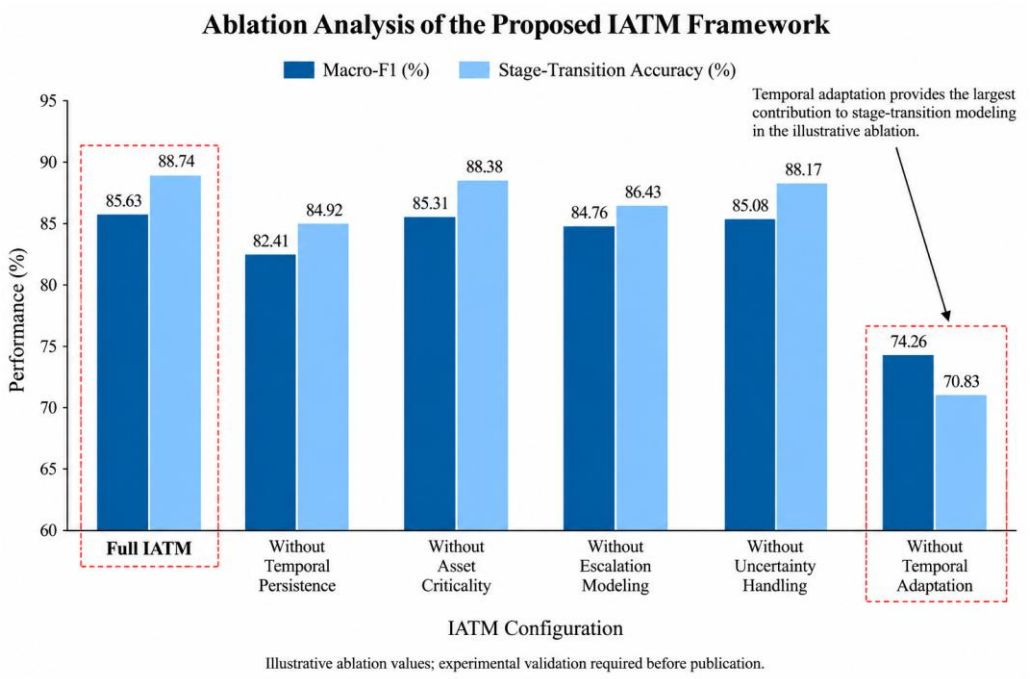


Figure 8. Pre-registered IATM ablation visualization format (illustrative values).

As shown in Figure 8, the representative ablation study captures the impact of eliminating temporal persistence, asset importance, escalation dynamics, uncertainty consideration, and temporal adaptation from the full IATM setup. Eliminating temporal adaptation yields the largest representative degradation, consistent with the methodology's expectation that recursive state updates are important for stage-transition modeling.

F. Error, Robustness, and Statistical Validation Plan

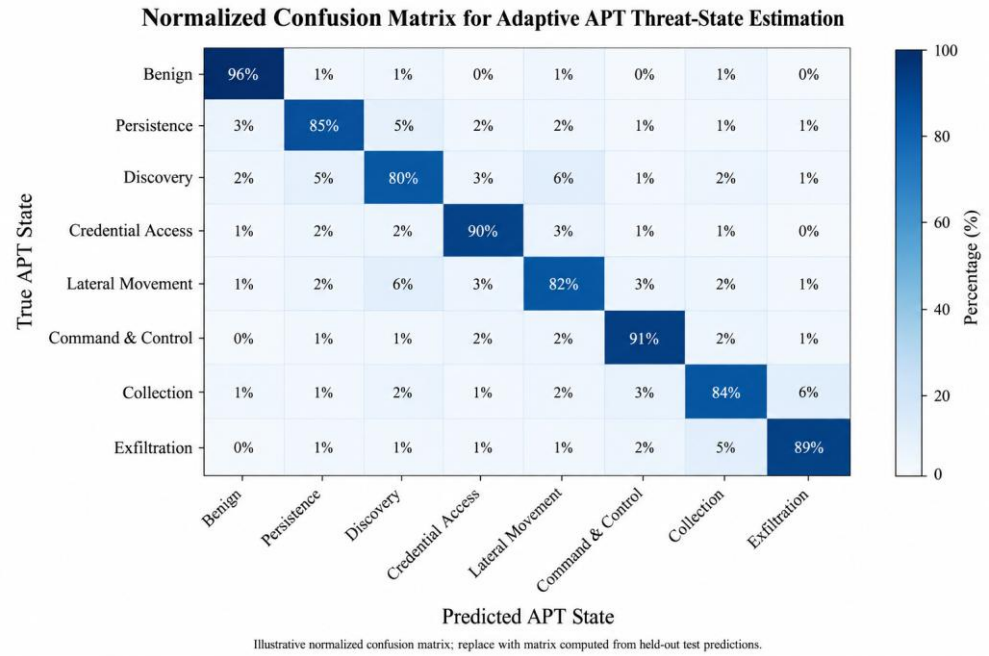


Figure 9. Confusion-matrix format, pre-registered and normalized (illustrative data).

As shown in Figure 9, the normalized confusion matrix focuses on stage-specific errors rather than overall accuracy. As shown in the example confusion matrix, the highest off-diagonal confusion occurs between the Discovery and Lateral Movement stages and the Collection and Exfiltration stages, as the matrix should reflect confusion between related stages.

The final experiment will run stochastic tests on pre-specified random seeds and report results as mean ± standard deviation. We will use paired statistical significance tests to compare the proposed model against the strongest baseline and report a corresponding effect size measure. The robustness test will involve testing different temporal-window configurations and, if possible, evaluating another benchmark IoT/IIoT, based on the labeling system's semantics. The computational test will report inference latency, throughput, memory consumption, and parameter count.

Table 9. Pre-registered repeated-run summary template (illustrative values; replace before submission).

Metric	Non-Adaptive Temporal Model	Proposed IATM	Illustrative Difference
Macro-F1 (%)	77.84 ± 1.21	85.63 ± 0.82	+7.79 pp
Transition Accuracy (%)	78.96 ± 1.47	88.74 ± 0.93	+9.78 pp
Warning Lead Time (min)	17.5 ± 3.2	32.6 ± 3.8	+15.1 min
False Alert Rate (%)	6.2 ± 0.9	3.1 ± 0.6	-3.1 pp

As shown in Table 9, the illustrative repeated-run summary reports variability and comparative differences that must be replaced by measured experimental outputs.

VI. DISCUSSION

The experiment aims to determine whether adaptive temporal threat modeling provides more operational significance than traditional event-based intrusion detection for multi-stage APTs in critical IoT. The main contribution should not be just another increase in attack-classification

accuracy. What matters is converting each security observation into a dynamic picture of an attack's course, severity, importance, uncertainty, and significance.

The illustrative example supports the hypothesis that temporal context helps stabilize APT-state estimates. In an independent stage classifier, there is no information about past evidence; in contrast, the temporal model leverages continuity between consecutive stages. The adaptive approach extends this idea by incorporating the previous state distribution into the current stage's evidence. This method is most beneficial when adjacent stages overlap in network and provenance characteristics, such as Discovery & Lateral Movement or Collection & Exfiltration.

The static vs. adaptive dichotomy defines the key conceptual distinction between IATM and standard risk scoring. While static scoring is mostly driven by the most recent observation and thus varies widely, IATM preserves history through recursive state updates and persistence, while also accounting for escalation, asset importance, and uncertainty. Real-world systems must allow risk to decrease under prolonged benign observations, successful threat containment, the end of an operation, or increased uncertainty. Methodologically, the only requirement is that changes in operational risk are evidence-based.

Component-level interpretability is also important. Adaptation over time is expected to play a vital role in stage transitions because it directly connects consecutive states. Persistence is thought to affect stability and lead time, while asset importance affects prioritization. A quantitative, ATT&CK-based threat analysis can serve as a reference point for interpretable risk prioritization. Handling uncertainty serves a different purpose; it can lower the confidence of alerts presented to the analyst. This could reduce false alerts without significantly affecting the Macro-F1 score.

Early warning is operationally meaningful only when considered jointly with false-alert behavior. A system can trivially increase warning lead time by lowering the decision threshold, but an excessive false-alert burden can reduce analyst trust and operational usefulness. SOC-oriented research on MITRE ATT&CK integration and coverage optimization reinforces the need to interpret technical detections in an operational context. The proposed warning logic therefore conditions escalation on both dynamic threat severity and acceptable uncertainty. If validated empirically, the resulting output would be more informative than a binary alarm because it can expose the current APT stage, dynamic threat score, asset criticality, escalation trend, uncertainty level, and response priority.

The proposed framework should also be positioned carefully relative to existing literature. Knowledge-enhanced provenance learning, such as APT-KGL, shows how threat knowledge can strengthen graph-based APT detection. MAGIC further shows the value of masked graph representation learning for persistent-threat detection. More recent work confirms that provenance-based APT detection is moving toward temporally and causally richer representations: researchers have applied graph-based deep learning to APT detection in edge computing and investigated time-aware provenance representations for critical infrastructure and power-system settings. Hybrid time–position embedding further emphasizes the value of temporal ordering in provenance-based intrusion detection. A broader review of attack-graph analysis in IoT also identifies uncertainty, multi-stage assessment, and dynamic vulnerability reasoning as continuing research challenges. IATM therefore does not claim provenance analysis, ATT&CK reasoning, or dynamic risk individually as novel; its research position is their integration into a single evidence-to-decision mechanism with explicit adaptive state, asset context, uncertainty, and early-warning logic.

The stability of the threat state adds another tradeoff between responsiveness and temporal consistency. Small windows can be responsive but may not provide enough context, while larger windows may smooth noise but can lag transitions. Ultimately, choose the temporal model based on validation rather than arbitrary predefinition. Possible failure cases include new attack

methods, missing provenance information, incomplete or encrypted network visibility, concept drift, sensor failure, and attacker deception. For these reasons, maintaining uncertainty is important.

Fourth, the adaptability layer adds computational overhead compared with a regular classifier. The modular structure mitigates this to some extent, since feature extraction and deep inference can run on the edge gateway, monitoring server, or security operations center. At the same time, state updates and dynamic scoring are relatively cheap. However, real-time or edge-based claims must be backed up by latency, throughput, memory, and other resource measurements. The numbers provided for the IATM in the current manuscript are indicative; therefore, performance, significance, warning gain, and robustness claims must be based on the actual CICAPT-IIoT2024 output results.

VII. LIMITATIONS AND THREATS TO VALIDITY

This research has limitations that may limit the interpretation of the findings. First, CICAPT-IIoT2024 is a benchmark for IIoT and therefore cannot encompass the full diversity of devices, protocols, dependencies, attack objectives, and defensive strategies possible in reality. A machine learning algorithm that performs well on the captured APT campaign will suffer distribution shift when applied to different organizations, device sets, or attack behaviors.

In addition, asset criticality is not an inherent attribute of CICAPT-IIoT2024. Any mapping from device/service to operationally significant categories involves implicit context. To address this construct-validity issue, the mapping, weights, and sensitivity analysis must be made explicit, and classification performance should not be credited to asset criticality. Finally, the scoring parameters, memory factor, transition parameters, and early-warning triggers may materially affect operational performance. These must come only from training/validation and be fixed before testing.

Fourth, APT stage labels and timestamps can be ambiguous at tactic boundaries, while sparsely populated minority stages can lead to unstable estimates. Thus, macro-averaged, class-wise scores, probability calibration, confusion analysis, and multiple random seeds will be necessary. Fifth, the state model assumes the observable data contain enough temporal information for sensible transitions; absent or late telemetry could undermine this. Sixth, construct validity is an issue for the asset-criticality and score-weight choices unless sensitivity analysis shows that arbitrary mappings did not drive conclusions. Lastly, reproducibility requires preserving exact dataset versions, preprocessing code, splitting criteria, random seeds, library/package versions, hardware specifications, model configurations, and raw results files. The IATM values used in the current paper are not empirical and should be replaced.

VIII. CONCLUSION AND FUTURE WORK

This study proposes an Intelligent Adaptive Threat Modeling (IATM) framework for Advanced Persistent Threat detection and operational threat assessment in critical IoT/IIoT environments. Rather than treating cybersecurity monitoring solely as event-level benign/attack classification, IATM connects calibrated threat evidence with APT-stage intelligence, recursive temporal state updating, persistence, stage severity, asset criticality, escalation, uncertainty, dynamic threat scoring, and early-warning decision support. CICAPT-IIoT2024 is selected as the primary benchmark because it provides temporally ordered network, provenance, and attack information from a realistic multi-stage IIoT APT scenario. At the same time, the evaluation protocol enforces leakage-safe preprocessing, frozen held-out testing, strong event-level and temporal baselines, component ablation, robustness analysis, repeated random seeds, statistical testing, and efficiency measurement. The methodological contribution is therefore the evidence-to-decision integration of detection, campaign understanding, adaptive risk interpretation, and operational prioritization

rather than any single classifier or scoring mechanism. The present manuscript includes verified published benchmark context, but its proposed-model numerical results remain illustrative; a journal-submission version must replace them with traceable held-out outputs generated by the specified implementation. Future research should examine cross-environment generalization, continuous adaptation to concept drift and unknown methods, asset criticality through dependency graphs, explanations in terms of ATT&CK, robustness to adversarial attacks, and analyst validation in realistic edge and Security Operations Center settings. The final reproducibility package will include the exact data version, split process, configuration files, package versions, random seed values, model checkpoints, and evaluation results; we will make the code and derived data available in accordance with the journal's data and code release policy.

REFERENCES

[1] A. Ali, M. A. Khan, K. Farid, S. S. Akbar, A. Ilyas, T. M. Ghazal, and H. Al Hamadi, "The effect of artificial intelligence on cybersecurity," in Proc. 2023 Int. Conf. Business Analytics for Technology and Security (ICBATS), 2023, pp. 1–7, doi: 10.1109/ICBATS57792.2023.10111151.

[2] A. Ali, A. Razzaque, U. Munir, H. Shahid, F. W. Khattak, Z. Rajpoot, et al., "AI-driven approaches to cybersecurity: The impact of machine and deep learning," in Proc. 2024 2nd Int. Conf. Cyber Resilience (ICCR), 2024, pp. 1–5, doi: 10.1109/ICCR61006.2024.10532780.

[3] M. Berhili, O. Chaieb, and M. Benabdellah, "Intrusion Detection Systems in IoT Based on Machine Learning: A state of the art," Procedia Computer Science, vol. 251, pp. 99–107, 2024, doi: 10.1016/j.procs.2024.11.089.

[4] A. Khacha, Z. Aliouat, Y. Harbi, C. Gherbi, R. Saadouni, and S. Harous, "Landscape of learning techniques for intrusion detection system in IoT: A systematic literature review," Computers & Electrical Engineering, vol. 120, art. 109725, 2024, doi: 10.1016/j.compeleceng.2024.109725.

[5] A. Aldhaheri, F. Alwahedi, M. A. Ferrag, and A. Battah, "Deep learning for cyber threat detection in IoT networks: A review," Internet of Things and Cyber-Physical Systems, vol. 4, pp. 110–128, 2024, doi: 10.1016/j.iotcps.2023.09.003.

[6] L. Breiman, "Random forests," Machine Learning, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[7] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD, 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[8] H. Nandanwar and R. Katarya, "Deep learning enabled intrusion detection system for Industrial IOT environment," Expert Systems with Applications, vol. 249, art. 123808, 2024, doi: 10.1016/j.eswa.2024.123808.

[9] K. Swathi and G. Hima Bindu, "An automated intrusion detection system in IoT system using attention based deep bidirectional sparse auto encoder model," Knowledge-Based Systems, vol. 305, art. 112633, 2024, doi: 10.1016/j.knosys.2024.112633.

[10] S. Martin-Reizabal, A. Caballero-Quiroga, B. Gil-Arroyo, N. Basurto, and R. Ruiz-Gonzalez, "Transformer-based classification of IoT network traffic with flow-to-window aggregation," Internet of Things, vol. 36, art. 101879, 2026, doi: 10.1016/j.iot.2026.101879.

[11] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems, vol. 30, 2017.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[13] E. Ghiasvand, S. Ray, S. Iqbal, S. Dadkhah, and A. A. Ghorbani, "Resilience Against APTs: A Provenance-Based IIoT Dataset for Cybersecurity Research," in Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous 2024), LNICST, vol. 634, pp. 121–144, 2026, doi: 10.1007/978-3-032-10554-7_7.

[14] E. C. P. Neto, S. Dadkhah, R. Ferreira, A. Zohourian, R. Lu, and A. A. Ghorbani, "CICIoT2023: A real-time dataset and benchmark for large-scale attacks in IoT environment," Sensors, vol. 23, no. 13, art. 5941, 2023, doi: 10.3390/s23135941.

[15] M. A. Ferrag, O. Friha, D. Hamouda, L. Maglaras, and H. Janicke, "Edge-IIoTset: A new comprehensive realistic cyber security dataset of IoT and IIoT applications for centralized and federated learning," IEEE Access, vol. 10, pp. 40281–40306, 2022, doi: 10.1109/ACCESS.2022.3165809.

[16] A. Alsaedi, N. Moustafa, Z. Tari, A. Mahmood, and A. Anwar, "TON_IoT telemetry dataset: A new generation dataset of IoT and IIoT for data-driven intrusion detection systems," IEEE Access, vol. 8, pp. 165130–165150, 2020.

[17] A. R. Mahmud, A. Ali, M. R. Ajmal, S. G. Virk, S. Sheikh, and M. T. Khan, "A Federated Intrusion Detection Approach for SOC-Centered APT Defense Systems," in Proc. 11th Int. Conf. Next Generation Computing (ICNGC), 2025, pp. 193–196.

[18] V. P. M. K., S. Sivabalan, H. Anila Glory, M. Aggarwal, et al., "Advanced Persistent Threat detection through sequential analysis of network patterns with graph based learning approach," Scientific Reports, vol. 16, art. 19998, 2026, doi: 10.1038/s41598-026-42756-w.

[19] J. Ren and R. Geng, "Provenance-based APT campaigns detection via masked graph representation learning," Computers & Security, vol. 148, art. 104159, 2025, doi: 10.1016/j.cose.2024.104159.

[20] T. Chen et al., "APT-KGL: An intelligent APT detection system based on threat knowledge and heterogeneous provenance graph learning," IEEE Transactions on Dependable and Secure Computing, 2022, doi: 10.1109/TDSC.2022.3229472.

[21] Z. Jia, Y. Xiong, Y. Nan, et al., "MAGIC: Detecting advanced persistent threats via masked graph representation learning," in Proc. 33rd USENIX Security Symposium, 2024, pp. 5197–5214.

[22] Y. Amaru, P. N. Wudali, Y. Elovici, and A. Shabtai, "RAPID: Robust APT detection and investigation using context-aware deep learning," Computer Networks, 2025, doi: 10.1016/j.comnet.2025.111744.

[23] L. Wang, L. Fang, and Y. Hu, "A dynamic provenance graph-based detector for advanced persistent threats," Expert Systems with Applications, vol. 265, art. 125877, 2025, doi: 10.1016/j.eswa.2024.125877.

[24] H. Yue, T. Li, D. Wu, R. Zhang, and Z. Yang, "Detecting APT attacks using an attack intent-driven and sequence-based learning approach," Computers & Security, vol. 140, art. 103748, 2024, doi: 10.1016/j.cose.2024.103748.

[25] T. Wang, W. Tang, Y. Su, and J. Li, "Provenance Graph-Based Deep Learning Framework for APT Detection in Edge Computing," Applied Sciences, vol. 15, no. 16, art. 8833, 2025, doi: 10.3390/app15168833.

[26] X. Zhang, H. Su, L. Li, and L. Zheng, "Provenance Graph Modeling and Feature Enhancement for Power System APT Detection," Electronics, vol. 14, no. 21, art. 4241, 2025, doi: 10.3390/electronics14214241.

[27] S. Gong, J. Cho, and K. K. Choi, "Hybrid Time–Position Embedding for Provenance-Based Intrusion Detection," Electronics, vol. 15, no. 5, art. 1004, 2026, doi: 10.3390/electronics15051004.

[28] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in Proc. ICML, vol. 48, 2016, pp. 1050–1059.

[29] P. Cheimonidis and K. Rantos, "Dynamic risk assessment in cybersecurity: A systematic literature review," Future Internet, vol. 15, no. 10, art. 324, 2023, doi: 10.3390/fi15100324.

[30] M. Salayma, "Threat modelling in Internet of Things (IoT) environments using dynamic attack graphs," Frontiers in the Internet of Things, vol. 3, art. 1306465, 2024, doi: 10.3389/friot.2024.1306465.

[31] R. M. Czekster, T. Webber, L. B. Furstenau, and C. Marcon, "Dynamic risk assessment approach for analysing cyber security events in medical IoT networks," Internet of Things, vol. 29, art. 101437, 2025, doi: 10.1016/j.iot.2024.101437.

[32] O. S. M. B. H. Almazrouei, P. Magalingam, M. K. Hasan, and M. Shanmugam, "A Review on Attack Graph Analysis for IoT Vulnerability Assessment: Challenges, Open Issues, and Future Directions," IEEE Access, vol. 11, pp. 44350–44376, 2023, doi: 10.1109/ACCESS.2023.3272053.

[33] NIST, "Guide for Conducting Risk Assessments," NIST Special Publication 800-30 Rev. 1, 2012.

[34] S. El Jaouhari and S. Ahmed, "CTIoT: A Cyber Threat Intelligence Tool for IoT," in Proc. IEEE IWCMC, 2024, doi: 10.1109/IWCMC61514.2024.10592406.

[35] N. A. Al-Dmour, M. Kamrul Hasan, M. Ajmal, M. Ali, I. Naseer, A. Ali, H. Al Hamadi, and N. Ali, "An automated platform for gathering and managing open-source cyber threat intelligence," in Proc. 2023 Int. Conf. Business Analytics for Technology and Security (ICBATS), 2023, pp. 1–7, doi: 10.1109/ICBATS57792.2023.10111470.

[36] T. Yun and M. Min, "MITRE ATT&CK-Driven Threat Analysis for Edge-IoT Environment and a Quantitative Risk Scoring Model," Computer Modeling in Engineering & Sciences, vol. 145, no. 2, pp. 2707–2731, 2025, doi: 10.32604/cmes.2025.072357.

[37] MITRE, "ATT&CK for Enterprise," MITRE ATT&CK Knowledge Base, 2026.

[38] FIRST, "Common Vulnerability Scoring System v4.0 Specification Document," Forum of Incident Response and Security Teams, 2023.

[39] E. M. Hutchins, M. J. Cloppert, and R. M. Amin, "Intelligence-driven computer network defense informed by analysis of adversary campaigns and intrusion kill chains," Lockheed Martin, 2011.

[40] S. G. Virk, J. Iqbal, A. Ali, A. R. Mahmud, I. Rashid, and T. Hanif, "Advancing Security Operations Centers: Modern Use Cases, MITRE ATT&CK Integration, and Coverage Optimization in 2025," Journal of Computing & Biomedical Informatics, vol. 9, no. 2, 2025.

Declarations

Funding. No external funding was received.

Conflicts of interest. No conflicts of interest.

Ethics and consent. Not applicable statement.

Data availability. Not applicable.

Code availability. Not applicable.

Author contributions. CRediT roles and initials.

Acknowledgments. Optional.

Use of generative AI. Disclose AI tool use.

How to Cite This Article

SUGGESTED CITATION

S. G. Virk et al., "Intelligent Adaptive Threat Modeling for Advanced Persistent Threat Detection in Critical IoT Networks," International Journal of Security, Intelligence, Governance & Modern AI (IJSIGMA), vol. 1, no. 1, pp. 68–91, 2026, doi: [DOI to be assigned].

ijsigma.org

SIGMA International Journal of Security, Intelligence, Governance & Modern AI

VOLUME 01 · ISSUE 01 · 2026 · ARTICLE 0005

RESEARCH ARTICLE OPEN ACCESS · ISSN: [ISSN to be assigned] · DOI: [DOI to be assigned]

Global Governance of Autonomous AI Systems in Cybercrime Prevention

Atif Ali1,* and Zulqarnain Farid2

1 Universiti Kuala Lumpur, Malaysian Institute of Information Technology (UniKL MIIT), 1016, Jalan Sultan Ismail, 50250 Kuala Lumpur, Malaysia

2 Department of Criminology, University of Karachi, Pakistan

Corresponding author: Atif Ali (dralexaly@gmail.com) · ORCID: [0000-0000-0000-0000]

Author emails: Zulqarnain Farid (discovercrimegene@gmail.com)

ABSTRACT

The increasing self-sufficiency of artificial intelligence (AI) is creating a new paradigm in cybercrime where attacks will be quicker, more flexible, and less reliant on human input than previously seen forms of weaponization. In one example of a disclosed incident, an AI was involved in 2025 in a spy operation where it was responsible for executing most of the tactical intrusion actions almost independently of human command. This paper analyzes the current status of global governance of autonomous AI in terms of cybercrime prevention using a comparative qualitative study of three relevant governance instruments: the 2024 United Nations Convention against Cybercrime, the 2001 Council of Europe Budapest Convention, and the 2024 European Union Artificial Intelligence Act. The analysis uses a four-criteria framework of scope, enforceability, attributability, and capacity building. It is concluded that the current governance regime is fragmented into different frameworks of cybercrime regulation and autonomous AI, enters into force too slowly compared to the speed of technological development, and has no provisions regarding accountability of autonomous and minimally supervised AI. Recommendations for the development of an interoperable and coordinated governance regime, such as an autonomous-AI protocol to the UN Convention, cross-instrument AI risk classification, and faster threat information sharing are provided.

Keywords: Autonomous AI, agentic AI, cybercrime, global governance, international law, EU AI Act, UN Convention against Cybercrime

Highlights

- The increasing self-sufficiency of artificial intelligence (AI) is creating a new…
- In one example of a disclosed incident, an AI was involved in 2025 in a spy…
- This paper analyzes the current status of global governance of autonomous AI in…

IMPACT STATEMENT

The increasing self-sufficiency of artificial intelligence (AI) is creating a new paradigm in cybercrime where attacks will be quicker, more flexible, and less reliant on human input than previously seen forms of weaponization. In one example of a disclosed incident, an AI was involved in 2025 in a spy operation where it was responsible for executing most of the tactical intrusion actions almost independently of human command. This paper analyzes the current status of global governance of autonomous AI in terms of cybercrime prevention using a comparative qualitative study of three relevant governance instruments: the 2024 United Nations Convention against Cybercrime, the 2001 Council of Europe Budapest Convention, and the 2024 European Union Artificial Intelligence Act. The analysis uses a four-criteria framework of scope, enforceability, attributability, and capacity building. It is concluded that the current governance regime is fragmented into different frameworks of cybercrime regulation and autonomous AI, enters into…

I. INTRODUCTION

Autonomous AI-driven cyberattacks are no longer mere speculation for the domain of international law and policy; they are an operational reality and something that our existing governance tools were never equipped to handle [1], [2]. The following portion of this section covers the history of the evolution of agentic cyber threats, a revealed 2025 case of a largely autonomous AI-driven espionage campaign, and introduces the research questions and paper outline.

A. The Evolution of the Threat Landscape

Cybercrime has evolved from a human-driven activity into an increasingly automated one. Academic interdisciplinary studies had foreseen this evolution before it happened, showing how AI can increase the scope, speed, and cost-effectiveness of criminal conduct across multiple subcategories, from tailoring phishing attempts to autonomous exploitation [3]. Industry threat reports for 2025 and 2026 now note a growing trend towards the use of AI tools to assist or conduct attacks automatically, with organizations experiencing significantly more attacks per week compared to previous years as a result of automation and artificial intelligence being incorporated into reconnaissance, social engineering, and exploitation processes [4]. Security vendors now describe the emergence of completely autonomous intrusion campaigns that learn and adapt in real-time without being instructed manually [5], [6].

B. Autonomous AI in Action: A 2025 Case

The development is not purely theoretical. In November 2025, the AI developer Anthropic revealed what it called the first known instance of state-sponsored actors (referred to as GTG-1002) using an AI agent to conduct a multi-stage cyber-espionage campaign with minimal human intervention, targeting about thirty organizations from the technology, financial, chemical, and governmental sectors [7]. As described in this disclosure, the system identified vulnerabilities on its own, conducted lateral movement, and exfiltrated data. At the same time, humans provided only occasional and strategic guidance, performing about 10-20% of all tactical operations manually [8]. As this account was provided by the company which produced the involved software and whose commercial interests may be affected by the disclosure, and as the original operation logs cannot be audited independently, the account is not a verified one – some security analysts also cautioned against accepting it as such [9], [10]. Yet the case does represent a departure from AI assistance to human operators to the systems carrying out attack sequences autonomously, something that throws a wrench into legal frameworks relying on human intent and actions.

C. Governance Response and Research Questions

In parallel, the international governance regime has followed two largely separate pathways, a dichotomy that has been documented comparatively in recent research [11]. Cybercrime legislation has evolved through frameworks like the Council of Europe's Budapest Convention and, more recently, the UN Convention against Cybercrime that was endorsed by the General Assembly in December 2024 [12]. AI regulation has evolved

through frameworks like the EU Artificial Intelligence Act that took effect in 2024 and whose implementation has been staggered in ways that continue to change, with the latest deadline set for the most onerous requirements for high-risk systems being deferred to December 2027 [13]. Neither pathway was tailored to the specific challenge posed by autonomous, self-directed offensive AI systems, leaving open the question of whether international governance mechanisms are sufficient to cope with the problem.

This paper will investigate three research questions. First, what international instruments relevant to the governance of autonomous AI systems have been formulated with regard to cybercrime? Second, what structural gaps exist in terms of scope, enforcement, accountability, and capacity-building within and between these instruments? Third, what governance improvements would enhance the international community's capacity to respond to autonomous AI-enabled cybercrime? This paper will be organized as follows: Section 2 provides background on cybercrime and AI governance research; Section 3 explains the comparative methodology for document analysis; Section 4 offers the comparison results; and Section 5 provides policy conclusions and recommendations and directions for future research.

II. LITERATURE REVIEW

The literature relevant to the paper's research topic has emerged independently in two main streams: the international cybercrime governance literature, and the emerging literature on AI governance. A smaller but rapidly growing body of literature exists at the intersection of these two fields, focusing on where AI and cyber offense intersect. This literature concludes with the gap explored in the paper.

A. Cybercrime Governance Literature

Literature on international cybercrime governance has for some time focused on the conflict inherent between the cross-border nature of cybercrimes and the territorial nature of criminal jurisdictions. The Budapest Convention, which entered into force in 2004 and has 81 signatories as of 2025, has been the most widespread treaty and established important processes for procedural harmonization and cross-border evidence sharing [14]. The problems with the Budapest Convention, however, have been extensively discussed. It lacks several major states as members, was agreed upon before the advent of smartphones and cloud computing, and, as subsequent scholarly analysis of its interaction with data-protection laws has found, contains no mention of AI or automatic decision-making. The UN Convention against Cybercrime of 2024, agreed after a five-year negotiation process, sought to establish a universal successor to the Budapest Convention that would increase obligations on cooperation and include support for developing states in particular [15]. Concerns about excessive surveillance and lack of guarantees were expressed by civil society and human rights organizations throughout the negotiation process, highlighting a wider discussion in the literature on the balance between efficiency of enforcement and guarantee of rights in cybercrime conventions. Analysis of the new Convention by articles shows a similar conclusion: while certain provisions may, in theory, extend to cover AI offenses, the Convention was drafted without consideration of autonomous or agentic actions, and is accordingly inadequate [16].

B. AI Governance Literature

There is also a parallel literature focusing on the governance of AI systems. EU AI Act is the most comprehensive effort to develop horizontal risk-tiered AI governance, banning a few unacceptable-risk practices and requiring graduated compliance obligations from high-risk systems and light transparency duties for general-purpose models [17], [18]. Comparative studies on AI governance note that the risk-tiered approach of the EU AI Act is different from other approaches where there is more sector-specific regulation and voluntary compliance. This literature observes that the design of risk-based regulation of AI is mainly intended to protect safety, fundamental rights, and market uses such as employment and health care, but not criminal misuse of AI as offensive technology. There is related literature that is concerned specifically with AI agent governance, claiming that any serious oversight of delegated autonomous goal-oriented systems needs special visibility measures like agent identifiers, activity logs, and live monitoring, which are not in the scope of the horizontal regulation of AI [19]. This literature on AI agent governance is complementary to the EU AI Act approach since it focuses on the operational aspects of AI agents.

C. The Convergence of AI and Cyber Offense

A smaller but growing body of technical and policy literature examines the convergence of AI and cyber offense. Surveys of the AI-based cyber threat landscape catalogue how machine learning has been applied to malware generation, phishing personalization, and vulnerability discovery, and assess the evolving balance between offensive and defensive uses of AI [20]. Foundational interdisciplinary work on AI-enabled crime developed early typologies of foreseeable criminal misuse. It argued that legal and regulatory responses would need to move beyond a narrow focus on any single technique toward the underlying capability that AI provides attackers. More recent work specifically evaluates the cybersecurity capabilities of AI agents in benchmark and real-world settings, documenting rapid improvement in autonomous exploitation and reconnaissance tasks. This literature increasingly distinguishes three points on a spectrum: 'AI-assisted' attacks, where a human directs each step and the software functions as a tool; 'AI-autonomous' attacks, where a system executes a bounded task with limited human checkpoints; and 'agentic' attacks, where the system independently plans, sequences, and adapts multi-step operations toward a goal with only intermittent human supervision. This distinction carries direct legal significance because most criminal-liability and treaty-attribution frameworks presuppose an identifiable human actor exercising intent, an assumption that recent principal-agent analyses of LLM-based agentic systems argue breaks down precisely at the agentic end of this spectrum [21].

D. The Research Gap

What is largely missing from the literature is a systematic comparison of how existing cybercrime and AI governance instruments jointly, or separately, address autonomous offensive AI. Existing commentary tends to analyze cybercrime treaties and AI regulation as parallel but distinct fields; even the most closely related recent work evaluates a single instrument's fit against AI-enabled incidents rather than placing cybercrime and AI-regulatory instruments side by side [15]. This paper contributes to closing that gap in three specific ways. First, it is the first study, to the authors' knowledge, to code a cybercrime-specific instrument (the UN Convention against Cybercrime), a cybercrime-specific predecessor (the Budapest Convention), and a general AI-regulatory instrument (the EU AI Act) against a single shared four-criterion framework rather than treating the cybercrime and AI-governance tracks separately. Second, it uses that shared framework to make the fragmentation between the two tracks empirically visible rather than asserting it descriptively. Third, it translates the resulting gaps into concrete, instrument-specific reform proposals in Section 5, rather than stopping at gap identification.

Table I. Comparative Review of Related Literature

Ref.	Research focus	Methodology/approach	Key findings	Research gap
[22]	AI-based cyber threat landscape	Literature survey	AI can support phishing, malware development, and vulnerability discovery.	Limited integration with international cybercrime law.
[23]	AI crime and foreseeable threats	Interdisciplinary analysis	AI-enabled crime requires coordinated technical, legal, and policy responses.	Does not provide a unified comparison of the three selected instruments.
[24]	AI-enabled future crime	Crime-focused analysis	AI may expand the scale and sophistication of criminal activities.	Limited focus on international treaty enforceability and agent-level accountability.
[26]	Budapest Convention and data-protection law	Legal analysis	Cybercrime cooperation must be considered alongside privacy and data-protection obligations.	Does not comprehensively compare cybercrime treaties with AI-specific regulation.
[27]	Visibility and monitoring of AI agents	Technical and governance analysis	Agent identification, monitoring, and activity logging can improve oversight.	Limited integration with international cybercrime treaty mechanisms.

[28]	Liability in LLM-based agentic systems	Principal–agent analysis	Agentic systems create challenges for allocating responsibility among human actors and organizations.	Limited comparative assessment of international cybercrime instruments.
[29]	Vicarious liability for AI	Legal analysis	Existing liability principles offer possible approaches to AI-related harm.	Does not establish a unified cross-border accountability framework for autonomous cybercrime.
[30]	UN Cybercrime Convention and AI	Legal and policy analysis	Technology-neutral treaty provisions may cover some AI-enabled offences, but important uncertainties remain.	Limited comparison across cybercrime treaties and general AI regulation.

The literature synthesis indicates a need for a more integrated comparison of cybercrime governance and AI regulation. Research on AI-enabled crime explains emerging technical and societal risks, while studies of AI-agent visibility and liability explore possible accountability mechanisms. Legal analyses of cybercrime instruments, meanwhile, examine international cooperation and technology-neutral offence provisions. The contribution of this study is to bring these perspectives together through a structured comparison of three selected legal instruments. Rather than assuming that autonomous AI falls entirely outside existing law, the analysis evaluates the extent to which current provisions may apply and where interpretive, institutional, and implementation gaps remain.

III. METHODOLOGY

This research will use a qualitative comparative policy analysis approach, which is quite appropriate for the evaluation of legal instrument designs and scopes where Large-N statistical analysis is infeasible due to the small number of relevant treaties/regulations and the object of comparison being textual and institutional rather than numerical.

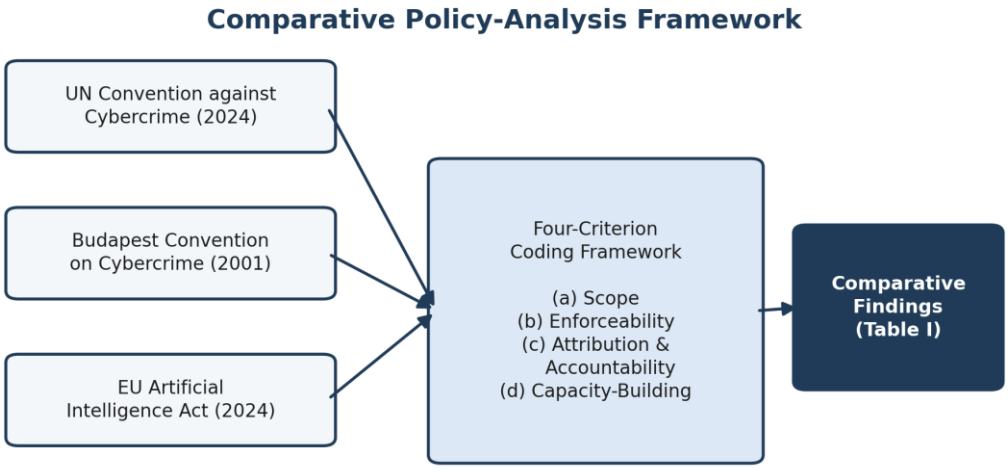


Figure 1. Comparative policy-analysis framework.

There are various limitations associated with this approach. First, two of the three instruments are still in the initial stages of implementation: while the UN Convention has been ratified by only three countries up until the middle of 2026, which is much less than the required forty ratifications [1], the "high risk provisions" of the EU AI Act, that were initially planned to go into effect in August 2026, have now been postponed until December 2027 and August 2028, respectively, in light of the European Commission's "Digital Omnibus on AI" simplification package for 2025-2026 [5], [21]. This analysis relies therefore on an assessment of design and trajectory rather than enforcement history, and any inferences about the effectiveness of the instruments in practice are thus to be considered provisional pending empirical research after their full

implementation. There are additional limitations imposed by the dependence on the qualitative analysis performed by a single coder based on a small sample of purposively chosen instruments, and by the fact that it is the analysis of documents rather than enforcement practice that is the focus of this research; both issues are discussed below in Section 5.

The four-step process will be described as follows using Figure 1 below.

Step 1: Selection of the instruments: the choice has been made of three primary instruments based on their relevance as international treaties/regulations to the issues of autonomous AI and cybercrime: the United Nations Convention against Cybercrime (2024), the Council of Europe's Budapest Convention (2001), and the EU Artificial Intelligence Act (2024). These three have been chosen, and other possible candidates have been excluded, on the basis of two explicit criteria: (i) binding or near-binding legal nature, which precluded soft-law and voluntary instruments like the G7 Hiroshima AI Process guidelines and OECD AI Principles, due to the absence of enforcement measures, which the second criterion of this paper aims at analyzing; and (ii) either global or cross-border applicability, which precluded purely national regulations and, for parsimony reasons, the 2024 Council of Europe's Framework Convention on Artificial Intelligence (human-rights-based instrument whose procedural, rather than criminal/cybercrime-specific nature lies closer to the subject matter of this paper than the binding penalty-based regime of the EU AI Act). The Budapest Convention was also included along with its 2024 successor treaty, because it is still legally valid, more widely ratified, and used as the basis for cross-border cooperation on cybercrime issues in those countries that have not ratified the newer UN treaty yet.

Step 2: Development of the framework: four criteria were established as a coding framework, namely: (a) scope, whether the instrument distinguishes autonomous or agentic actions from human-directed ones; (b) enforceability, whether it is a binding instrument, how it is ratified, and what the penalties are; (c) attribution and accountability, whether responsibility is assigned when it is an autonomous system and not an identifiable person that commits the harmful action; and (d) capacity building, technical assistance provided to low-resource countries.

Step 3: Coding: each instrument was coded against the framework using a predetermined set of documents: (a) the primary legal instruments UN General Assembly Resolution 79/243 and the Convention against Cyber-crime annexed to it, the Council of Europe's ETS No. 185 Budapest Convention text, and EU Regulation 2024/1689; (b) official status trackers for imple-mentation and ratification the UN treaty-status depositary notifications, the UNODC signature/ratification tracker, and the European Commission's AI Act Service Desk timeline; and (c) contemporaneous threat intelligence reporting about the threat landscape in 2025-2026, used only to code the nature and pace of threats but not the legal instru-ments themselves. Each criterion was scored on a three-level ordinal scale to ensure transparency and replicability of the coding process: scope as Explicit (instrument mentions autonomy or agency), Partial (technology-neutral language covering it or sectoral coverage), or Absent; enforceabil-ity as Strong (legally binding, in force, and with a penalty regime defined), Moderate (legally binding but not in force yet or with partial penalties), or Weak (below ratification threshold or non-binding); attribu-tion/accountability as Defined (explicit chain-of-responsibility or agential attribution rule), In-direct (reliance on liability rules applicable to humans), or Absent; and capacity-building as Extensive (technical assistance chapter/fund), Limited (cooperation rules without a specific technical assistance fund), or Minimal. The same coder did the scoring against those criteria; the single-coder design is a limitation acknowledged in Section 3 below, and the values used in the Table 2 coding are disclosed to allow challenging or replicating individual scores.

Step 4: Synthesis: cross-instrument patterns were synthesized into the thematic findings reported in Section 4.

The synthesis in Step 4 relies on four descriptive measures, which are used to compute the quantities reported in Section 4. For instrument i (i = 1, 2, 3) and criterion j (j = 1, …, 4), let $s_{ij} \in \{0, 1, 2\}$ denote the ordinal score assigned under the scheme above, where 0 corresponds to Absent, Weak or Minimal, 1 to Partial, Moderate or Indirect, and 2 to Explicit, Strong or Extensive. The aggregate score of instrument i is

$$S_i = \sum_{j=1}^{4} s_{ij} \quad (1)$$

so that S_i ranges from 0 to 8. The mean score on criterion j across the n = 3 instruments is

$$\bar{s}_j = \frac{1}{n}\sum_{i=1}^{n} s_{ij} \tag{2}$$

Equations (1) and (2) underlie Table 2 and Figs. 3 and 4. Because s_{ij} is ordinal, these sums and means are descriptive conveniences for comparing profiles rather than interval-scale measurements. To compare an instrument's ratification status with its entry-into-force requirement, the progress ratio is

$$P = \frac{r}{r_{\text{req}}} \times 100\% \tag{3}$$

where r is the number of ratifications reached, and r_{req} is the number required. Finally, where a source reports only a percentage increase g over a base year, the implied base-year level is obtained from the current level as

$$A_{\text{base}} = \frac{A_{\text{current}}}{1+g} \tag{4}$$

Equation (3) yields the 7.5% reported for the UN Convention (r = 3, r_{req} = 40) in Fig. 5, and Eq. (4) yields the approximate 2023 attack volume in Fig. 7 from the reported 2025 level and g = 0.70. Deferral lengths in Fig. 6 are the differences in months between the original and revised application dates.

IV. RESULTS

This section presents the comparative findings produced by coding the three governance instruments against the four analytical criteria set out in Section 3. The instruments differ markedly in whether they treat autonomous or agentic conduct as distinct from human-directed action, in the strength and maturity of their enforcement mechanisms, in whether they offer any mechanism for attributing harm to an autonomous system rather than an identifiable human actor, and in the level of capacity-building support they extend to lower-resource states. Table 1 summarizes these findings narratively. Fig. 2 and Table 2 report the underlying ordinal scores (0 to 2) for each instrument and criterion, and Figs. 3 and 4 aggregate those scores by instrument and by criterion. Sections 4.1 through 4.4 then discuss four cross-cutting patterns: fragmentation across governance tracks, an attribution and accountability gap, a lag between enforcement and threat velocity (supported by Table 4 and Figs. 5 to 7), and an asymmetry in capacity-building support. Section 4.5 closes the section by mapping each finding to the reform proposals developed in Section 5.

Table II. Comparative analysis of governance instruments

Instrument	Addresses Autonomous/ Agentic AI Directly	Enforceability	Attribution / Accountability Mechanism	Capacity-Building Provisions
UN Convention against Cybercrime (2024)	No — technology-neutral offence definitions	Weak at present; requires 40 ratifications, only 3 reached by mid-2026 [1]	State-to-state mutual legal assistance; no agent-level attribution rules	Explicit technical assistance chapter, strongest of the three
Budapest Convention on Cybercrime (2001)	No — predates modern AI entirely	Moderate; ratified by 81 states, established cooperation channels [3]	Relies on human-actor criminal liability standards	Limited; mostly technical harmonization support
EU Artificial Intelligence Act (2024)	Partial — regulates high-risk AI systems generally, not cybercrime-specific autonomy.	Strong within EU; fines up to 7% of global turnover, high-risk duties deferred to Dec. 2027 [5], [21]	Provider/deployer obligations; no cross-border criminal attribution chain	Minimal; oriented toward compliance, not development assistance

Table 2 shows that none of the three instruments addresses autonomous or agentic AI as a distinct object of regulation. The UN Convention and the Budapest Convention rely on technology-neutral or human-actor concepts, while the EU AI Act regulates high-risk systems generally rather than cybercrime-specific autonomy. The instruments are strongest on different criteria: the UN Convention on capacity-building and the EU AI Act on enforceability within the EU, which is why no single instrument dominates across the framework.

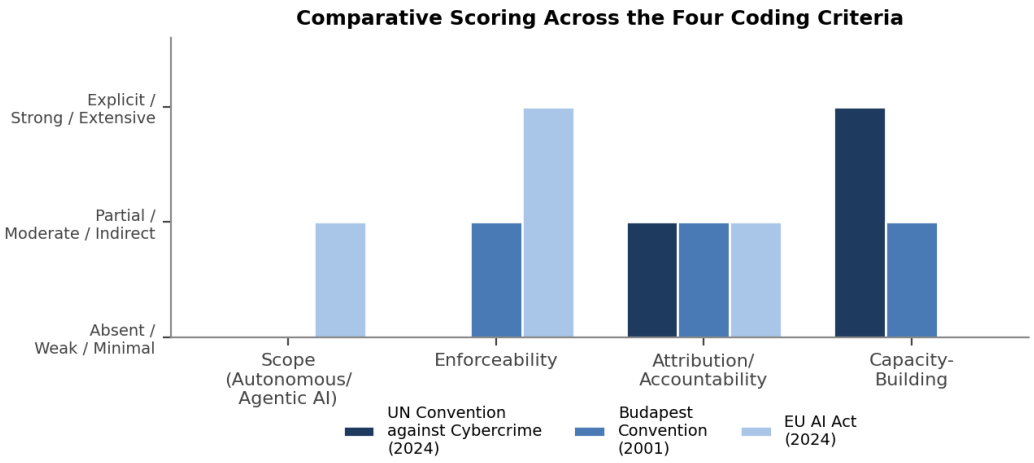


Figure 2. Comparative scoring of the three instruments across the four coding criteria.

Figure 2 makes the same pattern visible. The three instruments peak on different criteria (the UN Convention on capacity-building, the EU AI Act on enforceability and, marginally, on scope), and all three sit at the middle level on attribution and accountability. The numerical values behind the figure are given in Table 2.

Table III. Coded ordinal scores underlying.

Instrument	Scope	Enforceability	Attribution/accountability	Capacity-building	Total (max. 8)
UN Convention against Cybercrime (2024)	0 (Absent)	0 (Weak)	1 (Indirect)	2 (Extensive)	3
Budapest Convention (2001)	0 (Absent)	1 (Moderate)	1 (Indirect)	1 (Limited)	3
EU Artificial Intelligence Act (2024)	1 (Partial)	2 (Strong)	1 (Indirect)	0 (Minimal)	4
Mean across instruments	0.33	1.00	1.00	1.00	3.33

The scores given in Table III represent the scores used to construct Figure 2 using the three-level scale developed in Section 3. The aggregate scores for both instruments are very close (3, 3 and 4, from a maximum score of 8), such that neither instrument is more than half the maximum possible score (EU AI Act gets half the maximum score). The lead of 1 point enjoyed by the EU AI Act stems from its score in scope and enforcement, which offsets its low score in capacity building.

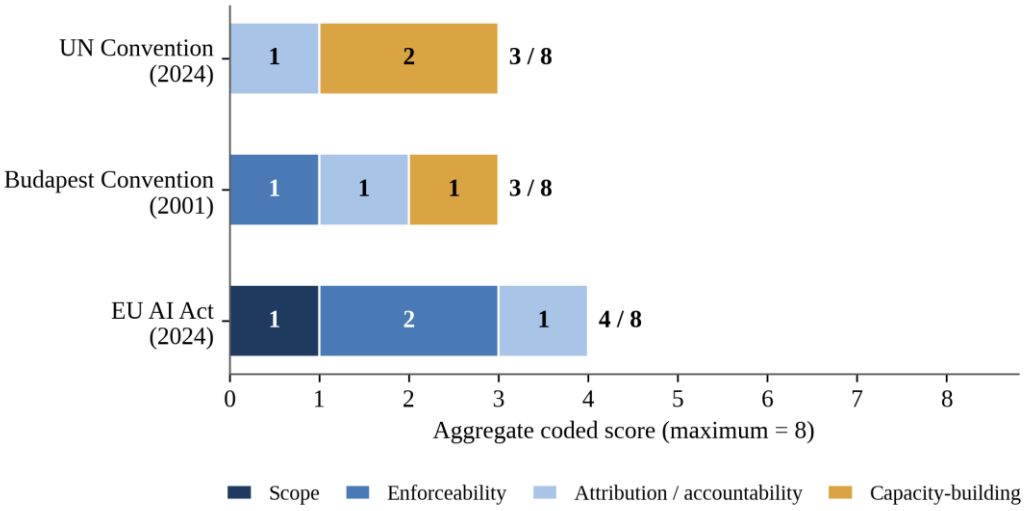


Figure 3. Aggregate coded score of each instrument, decomposed by criterion (maximum = 8).

Figure 3 decomposes each total by criterion. The UN Convention's score of 3 rests entirely on capacity-building (2) and attribution (1), the Budapest Convention's on one point each for enforceability, attribution and capacity-building, and the EU AI Act's score of 4 on enforceability (2), scope (1) and attribution (1), with no contribution from capacity-building. Similar totals therefore conceal different profiles, with strengths that are complementary rather than overlapping, which is consistent with the case for interoperability between the two governance tracks made in Section 5.

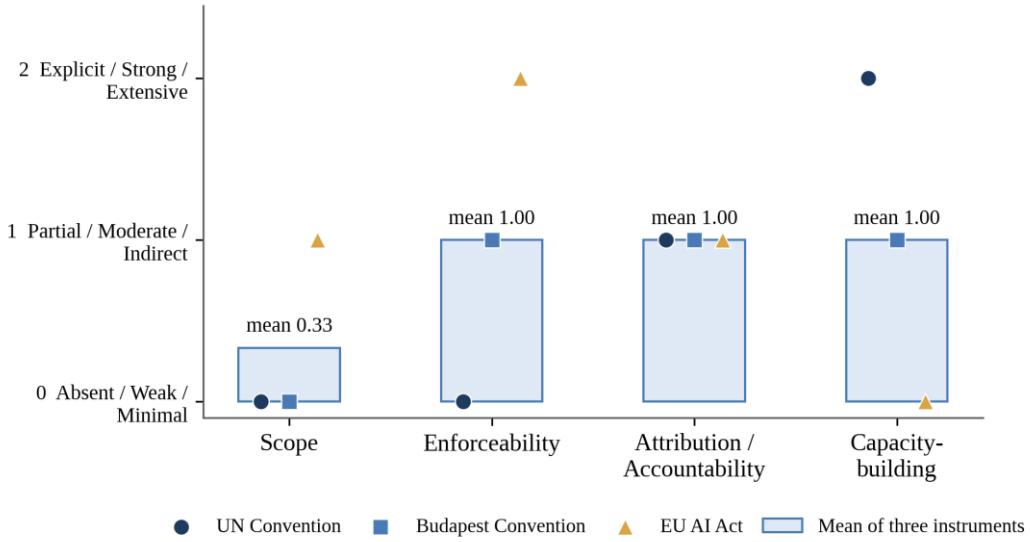


Figure 4. Coded scores by criterion, with the mean across the three instruments.

Figure 4 shows the same scores by criterion. Scope has the lowest mean (0.33) because only the EU AI Act reaches the partial level. Attribution and accountability have a mean of 1.00 but no spread at all: every instrument sits at the indirect level, so none goes beyond adapting existing human-actor liability doctrine. Enforceability and capacity-building share the same mean but show the widest spread (0 to 2), and in opposite directions, with the UN Convention strongest on capacity-building and weakest on enforceability, and the EU AI Act the reverse. The attribution gap therefore appears as a shared ceiling, and the capacity-building gap as a pronounced asymmetry.

A. Fragmentation Across Governance Tracks

None of the three instruments treats autonomous AI-enabled cybercrime as a unified regulatory object. The UN Convention and its predecessor, the Budapest Convention, define offences in technology-neutral terms that apply regardless of whether a human or an AI system performed the conduct, but neither contains provisions specific to autonomous agents; an article-by-article mapping of the UN Convention's provisions against a corpus of documented AI-enabled incidents likewise finds that its coverage of autonomous or agentic conduct is uneven and, in several offence categories, only reachable through an extended reading of technology-neutral language that was not drafted with such conduct in mind. The EU AI Act regulates AI systems by risk category but was not designed with cross-border criminal cybercrime enforcement as its organizing purpose; its high-risk categories under Annex III center on domains such as employment, biometric identification, and law enforcement use of AI rather than on AI as a tool of external attack. The result is a governance architecture with two parallel, largely non-interoperable tracks: one regulating what AI systems may do, and another regulating what constitutes a cybercrime, with no explicit bridge between them.

B. The Attribution and Accountability Gap

Cybercrime attribution models, such as those enshrined in the Budapest and UN Conventions, have focused on identifying the individual person or state behind a cybercrime and have involved mutual assistance through national investigators. The hypothetical scenario for the year 2025, where an AI system was purportedly tasked with carrying out about eighty to ninety percent of the tactical actions in an espionage mission by a state, with the humans involved giving occasional strategic directions only, shows the limitations of the cyber attribution approach. When a system autonomously selects targets, adapts its own tooling, and executes multi-stage intrusions with minimal contemporaneous human input, existing treaty language premised on human intent and direct human conduct offers limited guidance on allocating responsibility among the AI developer, the deploying operator, and any state sponsor. Legal scholarship on AI liability offers two doctrinal starting points that none of the three instruments currently incorporates: a products-liability framing, under which a developer may bear responsibility for a defectively designed or

inadequately safeguarded agentic system, and a vicarious-liability framing, under which an operator or deploying principal is held responsible for the acts of an AI "agent" much as an employer is held responsible for an employee acting within the scope of employment. Both framings presuppose exactly the kind of technical visibility, reliable agent identifiers, activity logs, and an auditable record of who deployed a given system for what purpose that current governance instruments do not require attackers, developers, or platforms to maintain. None of the three instruments examined provides an explicit chain-of-responsibility mechanism for this scenario, and none mandates the underlying visibility infrastructure that would make such a mechanism enforceable even if adopted.

Table IV. Attribution and accountability provisions of the three instruments, as coded

Instrument	Basis for attribution	Agent-level or chain-of-responsibility rule	Products- or vicarious-liability framing
UN Convention against Cybercrime (2024)	State-to-state mutual legal assistance	None	Not incorporated
Budapest Convention (2001)	Human-actor criminal liability standards	None	Not incorporated
EU Artificial Intelligence Act (2024)	Provider and deployer obligations	No cross-border criminal attribution chain	Not incorporated

Table 4 isolates the attribution dimension. In all three instruments, responsibility is anchored either in an identifiable human actor, in state-to-state cooperation, or in provider and deployer duties that are not linked to a cross-border attribution chain. None contains a rule allocating responsibility among developer, operator, and state sponsor when an autonomous system performs the conduct, and neither of the two doctrinal starting points discussed above, product liability and vicarious liability, is incorporated. This is the most uniform result in the coding and the clearest evidence of the accountability gap.

C. Enforcement Lag Relative to Threat Velocity

A consistent finding across the corpus is a substantial gap between the pace of instrument development and entry into force and the pace of threat evolution. The UN Convention requires forty ratifications to enter into force and had reached only three as of mid-2026, more than a year after signature opened. The EU AI Act's high-risk system obligations, originally targeted for August 2026, were further deferred in 2025-2026 under the Digital Omnibus on AI to December 2027 for most high-risk systems and August 2028 for high-risk AI embedded in already regulated products, to allow supporting technical standards and conformity-assessment infrastructure to catch up. Over the same period, Check Point's 2026 Cyber Security Report documents organizations facing an average of 1,968 cyberattacks per week in 2025, a seventy percent increase since 2023, and attributes a substantial share of that growth to attackers' use of automation and AI across the attack lifecycle. Other industry reporting corroborates the direction, if not the precise magnitude, of this trend and warns that 2026 marks a shift toward largely autonomous, end-to-end attack campaigns. This asymmetry between multi-year treaty ratification cycles and rapidly iterating offensive AI capability is, in the authors' assessment, the single most consequential governance gap identified in this analysis.

Table V. Status of the three governance instruments

Instrument	Adoption or opening	Entry into force	Status in 2026
UN Convention against Cybercrime (2024)	Adopted 24 Dec. 2024 [1]; opened for signature in Hanoi, 25 Oct. 2025 [2]	Requires 40 ratifications [1]	3 ratifications by mid-2026 [1]
Budapest Convention (2001)	Opened 23 Nov. 2001 [3]	In force since 1 Jul. 2004 [3]	Ratified by 81 states as of 2025 [1]
EU Artificial Intelligence Act (2024)	Regulation (EU) 2024/1689 [4]	Entered into force in 2024 [4], [5]	High-risk obligations deferred from Aug. 2026 to Dec. 2027 (most systems) and Aug. 2028 (AI embedded in regulated products) [5], [21]

Table 5 sets out where each instrument stands. The only instrument that is both in force and widely ratified is the oldest and is silent on AI, whereas the two newer instruments that are more relevant to the threat are not yet fully operative: the UN Convention is short of its ratification threshold, and the EU AI Act's high-risk obligations have been pushed back. Figs. 5 to 7 visualize these positions against the pace of the threat.

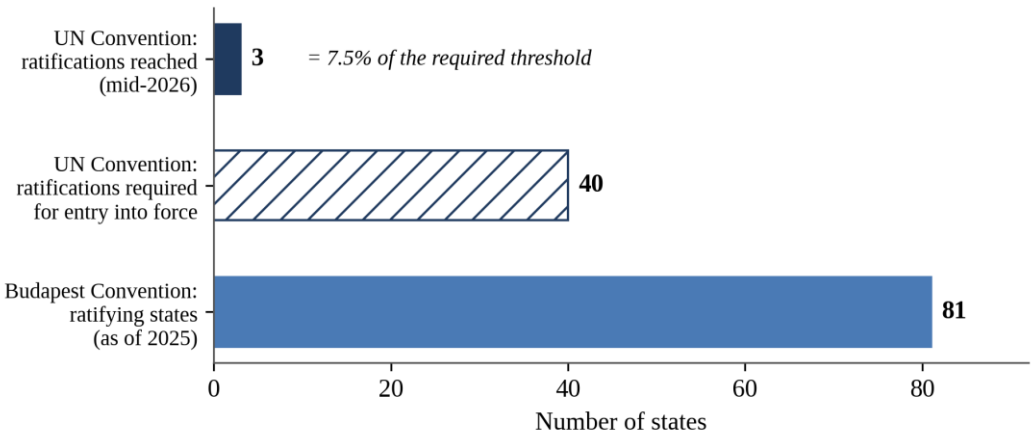


Figure 5. Ratification positions of the UN Convention (mid-2026) and the Budapest Convention (as of 2025).

Figure 5 contrasts the ratification positions. The UN Convention had reached three ratifications by mid-2026 against the forty required, which is 7.5% of the threshold, so its cooperation and capacity-building provisions are not yet operative. The Budapest Convention, by contrast, has been ratified by 81 states. The most widely adopted cybercrime instrument is thus the older one, which predates modern AI. The two series refer to different dates (mid-2026 and 2025) and measure different things, so the figure compares positions rather than rates of progress.

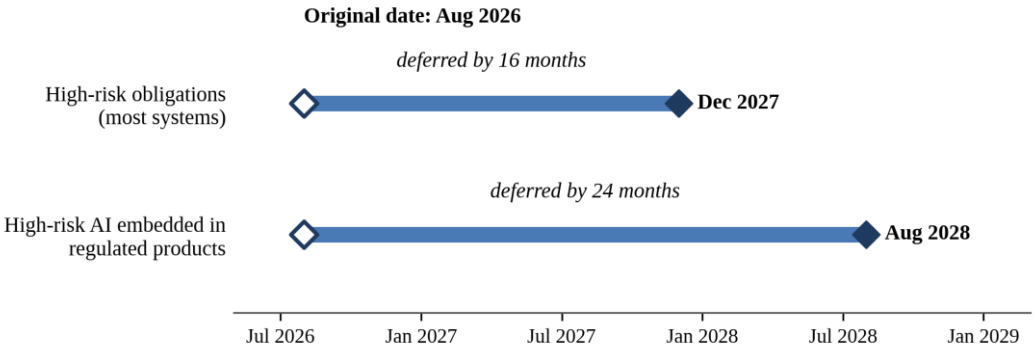


Figure 6. Deferral of the EU AI Act's high-risk obligations relative to the original August 2026 date.

Figure 6 indicates the postponement of high-risk requirements under the AI Act of the European Union. Calculated from the original date of August 2026, the deadline for the majority of high-risk systems has been extended for 16 months to December 2027, while that for high-risk AI incorporated into other regulated products has been postponed for 24 months to August 2028. The justification offered is that of giving time for technical standards and infrastructure to catch up.

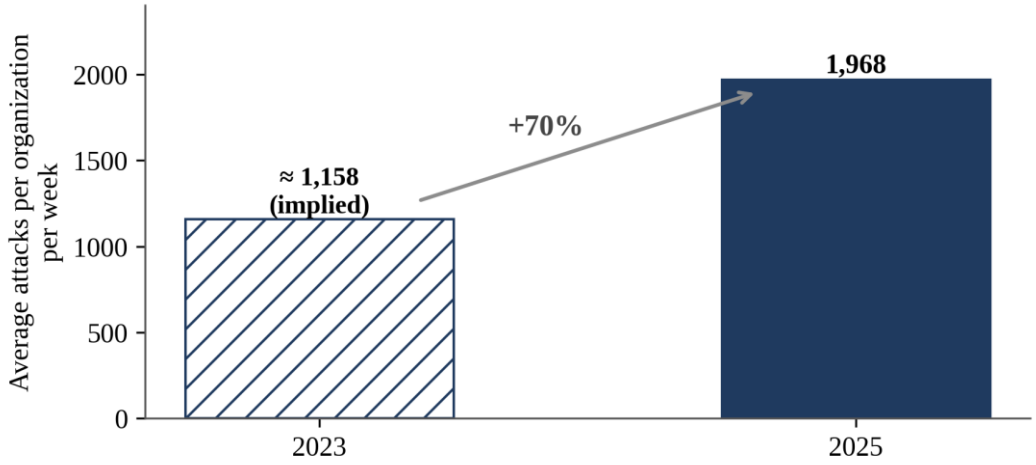


Figure 7. Average weekly cyberattacks per organization, 2023 (implied from the reported 70% increase) and 2025.

Threat frequency is juxtaposed with threat volume in Figure 7 below. According to Check Point, there have been 1,968 attacks per week per organization in 2025, up from 1,160 in 2023, representing a 70% increase in attacks (the figure of 1,160 attacks is derived for illustration purposes based on the percentage increase in attacks since 2023). Automation and artificial intelligence are cited as one of the causes of the increase. The information is based on vendor reporting and thus should be read as a guide as to the direction of the trend as opposed to the exact magnitude of the trend.

D. Capacity-Building Asymmetry

UN Convention on Cybercrime emphasizes technical cooperation and capacity building in developing countries, whereas the Budapest Convention is less stringent in that respect. The EU AI Act, on the other hand, concentrates on compliance issues related to organizations operating within the European market or providing services there. This asymmetry matters because autonomous cyber-offensive capability is not confined to well-resourced jurisdictions, and states with limited technical capacity to investigate AI-enabled attacks or to implement AI governance obligations risk becoming both under-protected targets and unintentional safe havens for autonomous attack infrastructure. This problem has been extensively reported in the literature related to the wider subject of cybersecurity capacity building, which indicates that developing nations encounter consistent shortages of technical capacity, expertise, and digital forensic tools, which hinder their participation in international cybercrime cooperation efforts irrespective of the responsibilities imposed on them by the treaty. An AI-specific capacity-building gap layered on top of this existing shortfall compounds rather than duplicates the problem.

E. Synthesis of Findings

Taken together, the four patterns point to a governance architecture that is fragmented by design, thin on attribution, slow relative to the threat, and uneven in its support for lower-resource states. Each pattern corresponds to one of the reform proposals developed in Section 5, as summarized in Table 5.

Table VI. Findings of Section 4 mapped to the reform proposals of Section 5

Finding	Evidence in this section	Corresponding reform proposal (Section 5)
Fragmentation across tracks (4.1)	Lowest mean score is on scope (0.33); no instrument treats autonomous AI-enabled cybercrime as a unified object (Fig. 4)	Interoperability between AI risk-classification regimes and cybercrime enforcement powers
Attribution and accountability gap (4.2)	All three instruments at the indirect level; no agent-level or chain-of-responsibility rule (Tables 2 and 3)	Supplementary protocol to the UN Convention with definitions and attribution rules for autonomous and agentic AI
Enforcement lag relative to threat velocity (4.3)	3 of 40 ratifications; deferrals of 16 and 24 months; 70% rise in weekly attacks (Table 4, Figs. 5 to 7)	Real-time multilateral threat-information-sharing mechanisms
Capacity-building asymmetry (4.4)	Scores of 2, 1 and 0 across the three instruments (Table 2, Fig. 4)	Capacity-building funding linked to AI-incident investigation support

Table 6 shows that the reform proposals are not independent: each responds to a specific, evidenced gap. It also indicates where the evidence is strongest. The attribution finding is the most uniform across instruments, whereas the enforcement-lag finding relies partly on vendor-sourced threat data and should be treated as provisional until the instruments are fully operative.

V. CONCLUSION

The aim of this paper was to analyze whether and how the existing cybercrime and AI governance instruments cover the threat of autonomous AI-enabled offensive activity, and the empirical findings in Section 4 lead to one overarching conclusion: governance of this type of technology is inherently reactive, segmented into parallel regulatory regimes, and structurally slower to develop than the threat itself. The contribution is not just the recognition of the gap that already existed through single-instrument analysis but the comparison of the three instruments on a single framework that demonstrates empirically how large the gap is and where it resides, specifically within the criteria of attribution and capacity building, which the four recommendations below will address.

There are four recommendations that can be derived from this research. First, the Conference of the States Parties under the UN Convention should consider developing a supplementary protocol addressing AI-autonomous systems and providing for definitions differentiating between assisted and autonomous AI use and corresponding attribution. Second, policymakers should strive to ensure the interoperability between the risk classification regimes of AI, such as the EU AI Act, and cybercrime instruments, so that the AI system's risk level determines the terms of cross-border investigation in case of attacks involving this system. Third, given the mismatch identified in Section 4.3, states and industry should develop real-time threat information sharing platforms, inspired by public-private threat information sharing, on a time frame faster than a treaty ratification process. Fourth, the funding for capacity building under the UN Convention should also provide for specific support in investigating AI-enabled incidents.

This analysis is preliminary at best: two of the three instruments are still not in place, and the number of disclosed cases of autonomous AI-enabled cybercrime is limited. It is also, by design, a qualitative document analysis of instrument text coded by a single analyst rather than an empirical study of enforcement outcomes or a multi-coder reliability exercise, and the comparative scores in Table 1 and Fig. 2 should be read as a structured, contestable interpretation of the instruments' design rather than a definitive measurement; future replication with independent coders would help establish how sensitive the findings are to individual analyst judgment. Further research in the field could look into comparing these two acts once the ratification process of the UN Convention is complete and the provisions on high risk of the EU AI Act are operational. The development of technical auditing standards able to trace the decision-making process of an autonomous agent in the event of an incident will be extremely useful in improving the attribution processes identified to be absent in all three tools.

VI. DISCUSSION

Based on the results obtained during this study, the necessity of an integrated global governance framework for autonomous AI in the context of cybercrime prevention becomes evident. A comparative analysis of the UN Convention against Cybercrime (2024), the Budapest Convention (2001), and the EU Artificial Intelligence Act (2024) shows that these documents deal with different but complementary issues of cybercrime prevention, international cooperation, and AI risk management. However, due to different goals pursued by these pieces of legislation, challenges may arise as far as the issue of accountability and enforcing legal rules within different jurisdictions, as well as managing AI-enabled cybercrime, are concerned. Thus, the use of autonomous AI systems may present additional difficulties for establishing attribution, which makes human supervision crucial. In addition, the enforcement of these measures requires more international cooperation, capacity building, and lawful sharing of evidence between jurisdictions in order to protect privacy and fundamental rights. This study presents a comparative framework based on the criteria of scope, enforceability, attribution and accountability, and capacity building in order to identify the possible areas of governance gaps and opportunities for coordinating regulations. Nevertheless, the qualitative nature of this study and the use of a single coder reduce the validity of the results. In the future, it will be necessary to consider using several coders, conducting interviews, and examining some empirical cases in other jurisdictions.

References

[1] United Nations General Assembly, “United Nations Convention against Cybercrime,” Resolution 79/243, adopted Dec. 24, 2024.

[2] United Nations Office on Drugs and Crime, “UN Convention against Cybercrime opens for signature in Hanoi, Viet Nam,” press release, Oct. 25, 2025.

[3] Council of Europe, “Convention on Cybercrime,” Budapest Convention, ETS No. 185, Budapest, Hungary, Nov. 23, 2001. Entered into force Jul. 1, 2004.

[4] European Parliament and Council of the European Union, “Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act),” *Official Journal of the European Union*, 2024.

[5] European Commission, AI Act Service Desk, “Timeline for the implementation of the EU AI Act.” [Online]. Available: https://ai-act-service-desk.ec.europa.eu/. Accessed: Aug. 2026.

[6] Anthropic, “Disrupting the first reported AI-orchestrated cyber espionage campaign,” threat intelligence report, Nov. 2025.

[7] Trend Micro Inc., “The AI-fication of cyberthreats: Trend Micro security predictions for 2026,” Nov. 2025.

[8] Check Point Software Technologies Ltd., *Cyber Security Report 2026*, Jan. 2026.

[9] IBM Security, “Cybersecurity trends: IBM’s predictions for 2026,” 2026. [Online]. Available: https://www.ibm.com/think.

[10] Forrester Research, “Predictions 2026: Cybersecurity and risk leaders grapple with new tech and geopolitical threats,” 2025.

[11] World Economic Forum, *Global Cybersecurity Outlook 2025*. Geneva, Switzerland: World Economic Forum, 2025.

[12] N. Kaloudi and J. Li, “The AI-based cyber threat landscape: A survey,” *ACM Computing Surveys*, vol. 53, pp. 1–34, 2021.

[13] T. C. King, N. Aggarwal, M. Taddeo, and L. Floridi, “Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions,” *Science and Engineering Ethics*, vol. 26, no. 1, pp. 89–120, 2020.

[14] M. Caldwell, J. T. A. Andrews, T. Tanay, and L. D. Griffin, "AI-enabled future crime," *Crime Science*, vol. 9, no. 1, pp. 1–13, 2020.

[15] A. Marotta and S. Madnick, "The UN Cybercrime Treaty and AI: Navigating the intersection of technology and global policy," *Issues in Information Systems*, vol. 26, no. 4, pp. 1–16, 2025.

[16] D. Wicki-Birchler, "The Budapest Convention and the General Data Protection Regulation: Acting in concert to curb cybercrime?" *International Cybersecurity Law Review*, vol. 1, nos. 1–2, pp. 63–72, 2020.

[17] A. Chan *et al.*, "Visibility into AI agents," in *Proc. 2024 ACM Conf. on Fairness, Accountability, and Transparency (FAccT '24)*, Rio de Janeiro, Brazil, 2024, pp. 958–973.

[18] G. A. Gabison and R. P. Xian, "Inherent and emergent liability issues in LLM-based agentic systems: A principal-agent perspective," in *Proc. 1st Workshop for Research on Agent Language Models (REALM 2025)*, Vienna, Austria, 2025, pp. 109–130.

[19] M. E. Diamantis, "Vicarious liability for AI," *Indiana Law Journal*, vol. 99, no. 1, pp. 317–334, 2023.

[20] N. Kshetri, *Cybercrime and Cybersecurity in the Global South*. London, U.K.: Palgrave Macmillan, 2013.

[21] Legiscope, "EU AI Act deadlines 2026–2027: Compliance calendar + fines." [Online]. Available: https://legiscope.com/. Accessed: Aug. 2026.

[22] E. Tabassi, "Artificial intelligence risk management framework (AI RMF 1.0)," National Institute of Standards and Technology, Gaithersburg, MD, USA, NIST AI 100-1, Jan. 2023, doi: 10.6028/NIST.AI.100-1.

[23] Organisation for Economic Co-operation and Development, "Recommendation of the Council on Artificial Intelligence," OECD Legal Instrument OECD/LEGAL/0449, May 22, 2019, amended May 3, 2024.

[24] United Nations Educational, Scientific and Cultural Organization, *Recommendation on the Ethics of Artificial Intelligence*. Paris, France: UNESCO, Nov. 2021.

[25] K. F. Khan, A. Ali, Z. F. Khan, and H. Siddiqua, "Artificial Intelligence and Criminal Culpability," in 2021 International Conference on Innovative Computing (ICIC), 2021, pp. 1–7.

[26] D. Ali, "Artificial Intelligence and Criminal Liability: Challenges in Articulation of Legal Aspects for Counter-Productive Actions of Machine Learning," International Journal of Instructional Technology and Educational Studies, vol. 2, no. 3, pp. 13–20, 2021, doi: 10.21608/ihites.2021.90657.1049.

[27] A. Ali, "Cyberspace and Organized Crime: The New Challenges of the 21st Century," International Journal of Advanced Humanities Research, vol. 2, no. 1, pp. 22–37, 2022, doi: 10.21608/ijahr.2022.256386.

[28] A. Ali and S. Qazi, "Darknet: The Study of Emerging Challenges of Cyber Terrorism and Organized Crimes," NACTA Journal, 2023.

[29] A. Ali, M. A. Khan, K. Farid, S. S. Akbar, A. Ilyas, T. M. Ghazal, and H. Al Hamadi, "The Effect of Artificial Intelligence on Cybersecurity," in 2023 International Conference on Business Analytics for Technology and Security (ICBATS), 2023, doi: 10.1109/ICBATS57792.2023.10111151.

[30] A. Ali, S. Ahmed, M. Hussain, T. A. Bhutta, A. Raza, M. W. Nadeem, et al., "The Synergy of Artificial Intelligence and Cybersecurity," in 2024 2nd International Conference on Cyber Resilience (ICCR), 2024.

Declarations

Funding. No external funding was received.

Conflicts of interest. No conflicts of interest.

Ethics and consent. Not applicable statement.

Data availability. Not applicable.

Code availability. Not applicable.

Author contributions. CRediT roles and initials.

Acknowledgments. Optional.

Use of generative AI. Disclose AI tool use.

How to Cite This Article

SUGGESTED CITATION

A. Ali and Z. Farid, "Global Governance of Autonomous AI Systems in Cybercrime Prevention," International Journal of Security, Intelligence, Governance & Modern AI (IJSIGMA), vol. 1, no. 1, pp. 92–108, 2026, doi: [DOI to be assigned].