

Article

Cross-Assessment & Verification for Evaluation (CAVe) Framework for AI Risk and Compliance Assessment Using a Cross-Compliance Index (CCI)

Cheon-Ho Min , Dae-Geun Lee  and Jin Kwak * 

Department of Cyber Security, Ajou University, Suwon 16499, Gyeonggi-do, Republic of Korea; mcheonho@ajou.ac.kr (C.-H.M.); dleorms486@ajou.ac.kr (D.-G.L.)

* Correspondence: security@ajou.ac.kr

Abstract

This study addresses the challenge of evaluating artificial intelligence (AI) systems across heterogeneous regulatory frameworks. Although the NIST AI RMF, EU AI Act, and ISO/IEC 23894/42001 define important governance requirements, they do not provide a unified quantitative method. To bridge this gap, we propose the Cross-Assessment & Verification for Evaluation (CAVe) Framework, which maps shared regulatory requirements to four measurable indicators—accuracy, robustness, privacy, and fairness—and aggregates them into a Cross-Compliance Index (CCI) using normalization, thresholding, evidence penalties, and cross-framework weighting. Two validation scenarios demonstrate the applicability of the approach. The first scenario evaluates a Naïve Bayes-based spam classifier trained on the public UCI SMS Spam Collection dataset, representing a low-risk text-classification setting. The model achieved accuracy 0.9850, robustness 0.9945, fairness 0.9908, and privacy 0.9922, resulting in a CCI of 0.9741 (Pass). The second scenario examines a high-risk healthcare AI system using a CheXNet-style convolutional model evaluated on the MIMIC-CXR dataset. Diagnostic accuracy, distribution-shift robustness, group fairness (finding-specific group comparison), and privacy risk (membership-inference susceptibility) yielded 0.7680, 0.7974, 0.9070, and 0.7500 respectively. Under healthcare-oriented weighting and safety thresholds, the CCI was 0.5046 (Fail). These results show how identical evaluation principles produce different compliance outcomes depending on domain risk and regulatory priorities. Overall, CAVe provides a transparent, reproducible mechanism for aligning technical performance with regulatory expectations across diverse domains. Additional metric definitions and parameter settings are provided in the manuscript to support reproducibility, and future extensions will incorporate higher-level indicators such as transparency and human oversight.



Academic Editors: Leandros A. Maglaras, Domenico Ursino and Aryya Gangopadhyay

Received: 13 November 2025

Revised: 22 December 2025

Accepted: 9 January 2026

Published: 10 January 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: AI risk management; Cross-Assessment & Verification for Evaluation (CAVe); Cross-Compliance Index (CCI); EU AI Act; ISO/IEC 23894 and 42001; NIST AI RMF

1. Introduction

The rapid advancement of artificial intelligence (AI) technology has become a part of everyday life for not only professionals but also general users. While AI offers convenience to users and can lead to dramatic improvements in productivity and accuracy in industrial settings, it also raises numerous concerns. Biased learning and information errors caused by faulty datasets, privacy concerns, and hallucinations are among the most common AI

risks. To ensure the safe development and use of AI, countries and international standards organizations are proposing AI risk management systems and governance frameworks.

The regulatory landscape represents a significant shift towards accountability. Recent enforcements, such as the European Union Artificial Intelligence Act (EU AI Act) [1] and the publication of International Organization for Standardization and International Electrotechnical Commission (ISO/IEC) 42001 [2], emphasize the need for rigorous risk management across the entire AI lifecycle. Furthermore, industrial applications of AI are facing increasing scrutiny regarding supply chain security [3] and cybersecurity bottlenecks in automation [4], necessitating standardized frameworks that can bridge the gap between high-level governance principles and technical implementation.

Currently, the National Institute of Standards and Technology (NIST) in the United States has established the AI Risk Management Framework (AI RMF) 1.0 [5]. Concurrently, the European Union (EU) is implementing stricter controls through the AI Act, establishing a regulatory framework based on risk levels. Another global standard, ISO/IEC 23894 and 42001, provides management guidelines for AI systems [2,6]. However, these frameworks operate with distinct objectives and terminologies, leading to fragmented assessment criteria. Because AI systems are deployed globally rather than confined to specific jurisdictions, a unified framework is needed to ensure consistent risk evaluation and compliance monitoring.

Recent global AI governance initiatives emphasize accountability and risk-based regulation, yet they largely remain qualitative in nature. Analyses of the EU AI Act point out that, while the Act introduces a risk-tiered regulatory structure, it does not provide a concrete or unified method for quantitatively assessing technical compliance across heterogeneous AI systems [7,8]. Similarly, the NIST AI Risk Management Framework offers flexible, non-binding guidance, but intentionally avoids prescribing numerical scoring or threshold-based evaluation mechanisms [9]. ISO/IEC standards such as 23894 and 42001 further focus on management-system requirements and organizational processes, leaving open the question of how technical performance metrics should be operationalized and compared in practice [10,11].

To address this challenge, this study proposes the Cross-Assessment & Verification for Evaluation (CAVe) Framework (detailed in Section 3), which identifies and cross-verifies common control elements across the NIST, EU, and ISO frameworks. The proposed method maps core controls to shared indicators and quantifies them into a Cross-Compliance Index (CCI) (elaborated in Sections 3.2 and 3.3), enabling a harmonized evaluation of regulatory alignment and system trustworthiness.

The main technical contributions of this study are summarized as follows:

- Cross-framework indicator intersection and mapping. We systematically identify and map a common set of measurable indicators across the NIST AI Risk Management Framework, the EU AI Act, and ISO/IEC AI standards by analyzing their control objectives and technical requirements (Section 3).
- Policy-aware quantitative compliance scoring. We propose a unified quantitative evaluation mechanism that integrates metric normalization, framework-specific thresholds, and evidence-based penalty factors into a single CCI, enabling consistent comparison across heterogeneous regulatory frameworks (Sections 3.2 and 3.3).
- Tunable evaluation reflecting regulatory priorities. We introduce a weighting scheme that allows the evaluation outcome to be adjusted according to the regulatory philosophy or domain context. The effects of framework-specific weights are incorporated into the final CCI computation and algorithmic evaluation process (Section 3.4).
- Empirical validation of cross-framework behavior. We validate the proposed framework through controlled experiments that demonstrate the effects of metric variation,

threshold enforcement, and policy-driven weighting on the resulting CCI, confirming the interpretability and reproducibility of the evaluation model (Section 4).

This article is organized as follows. Section 2 reviews related work. Section 3 introduces the proposed CAVe framework. Section 4 presents the validation procedures and experimental results. Section 5 provides a detailed discussion, and Section 6 concludes the paper.

2. Related Work

This section reviews representative works across three major areas relevant to this study: (1) international AI governance frameworks such as NIST AI RMF, the EU AI Act, and ISO/IEC standards; (2) ontology- and mapping-based approaches for structuring AI risks; and (3) recent advances in AI security and privacy. The objective is to clarify how prior studies approach AI risk management and to highlight the lack of a quantitative, cross-framework evaluation mechanism. These limitations motivate the development of the CAVe framework.

2.1. NIST AI RMF

NIST released the AI Risk Management Framework (AI RMF 1.0) in 2023 as a recommended guideline to ensure the safe and accountable deployment of AI systems [5]. In 2024, NIST expanded it through the Generative AI Profile to respond to emerging risks from Generative AI [12].

Barrett et al. suggested practical measures for mitigating severe AI risks [13]. However, the AI RMF deliberately refrains from defining explicit numerical thresholds or aggregated compliance scores, instead prioritizing flexibility and context-awareness. As noted by Brundage et al. [9], this design choice strengthens adaptability but limits direct comparability and cross-jurisdictional compliance evaluation.

2.2. EU AI Act

The EU AI Act is the world's first comprehensive AI law, which entered into force in August 2024 [1]. It will apply from August 2026. It classifies AI systems into four risk levels—unacceptable, high, limited, and minimal—and imposes strict obligations on high-risk AI systems, including requirements for accuracy, robustness, data governance, security, and human oversight.

At the same time, Smith et al. highlight that although the Act introduces categories for GPAI and systemic risks, the criteria and documentation requirements remain ambiguous given the rapid evolution and diversity of AI models [14].

Several legal and policy-oriented studies have noted that the EU AI Act, despite its comprehensive scope, lacks an explicit quantitative compliance assessment mechanism that can be applied consistently across domains and model types. Veale and Zuiderveen Borge-sius [7] argue that this ambiguity complicates practical enforcement, while Hacker et al. [8] highlight similar challenges for general-purpose and foundation models. These observations motivate the need for a complementary technical framework that can translate regulatory intent into measurable indicators.

2.3. ISO/IEC 23894 and ISO/IEC 42001

The ISO/IEC has developed several international standards to support AI risk management and governance. Among them, ISO/IEC 23894 provides risk management guidance based on ISO 31000, outlining procedures for identifying, analyzing, and mitigating AI-related risks [6]. ISO/IEC 42001, derived from the management system structures of ISO 9001 and ISO/IEC 27001, specifies the requirements for establishing and maintaining an

Artificial Intelligence Management System (AIMS) [2]. Complementary standards—such as ISO/IEC 22989 (concepts and terminology), ISO/IEC 42006 (transparency), and the ISO/IEC 5259 series (data quality and audit)—further extend this ecosystem, forming a comprehensive foundation for AI governance [15].

Boza and Evgeniou emphasize that embedding organizational governance and accountability mechanisms into such frameworks is critical for strengthening responsible AI implementation across sectors [16].

While ISO/IEC 23894 and ISO/IEC 42001 establish structured management-system requirements, prior studies emphasize that such standards primarily address organizational processes rather than measurable technical performance. Morley et al. [10] and Schiff et al. [11] identify a persistent gap between high-level AI principles and their operationalization, particularly in terms of quantitative evaluation and auditability.

2.4. Related Works on AI Risk Management and Governance

Golpayegani et al. proposed the AI Risk Ontology (AIRO), which is grounded in the proposed EU Artificial Intelligence Act and the ISO 31000 risk-management standard [17]. Xia et al. compared and analysed 16 risk assessment frameworks and emphasised the need for a connected-to-aira (C2AIRA) approach [18]. Compared with ontology-based or mapping-oriented approaches such as AIRO and C2AIRA, which primarily categorize and define AI risks, CAVe differs in that it provides a quantitative evaluation mechanism. Prior works do not define a unified scoring function or cross-framework metric, and therefore cannot produce measurable or comparable compliance scores. In contrast, CAVe formalizes indicator selection, normalization, thresholding, evidence penalties, and framework-level weighting into a single computable CCI. This makes CAVe operational rather than descriptive, enabling reproducible and regulator-aligned assessments across heterogeneous AI governance standards.

Recently, Karras presented an ethical AI framework (FRE-AIDT) that includes transparency, fairness, and accountability [19]. Cui et al. analysed the AI governance structure of 6G networks [20].

These prior studies have collectively underscored the importance of AI risk governance, yet none have proposed a quantitative or integrative method for its evaluation. The CAVe framework of this study stands out in that it integrates and quantifies these approaches.

2.5. Recent Advances in AI Security and Privacy

In parallel with the development of AI governance frameworks, recent research has revealed that modern AI systems are increasingly exposed to sophisticated security and privacy threats. Beyond traditional concerns such as data quality, bias, or model misbehavior, attackers can now exploit vulnerabilities at both the model and physical levels to compromise system integrity.

Physical-domain backdoor attacks such as FIGhost demonstrate that adversaries can implant fluorescent-ink-based triggers into real-world objects, enabling stealthy and flexible manipulation of traffic-sign recognition models and other safety-critical perception systems [21]. Such attacks highlight that AI security risks extend beyond digital environments and can directly affect deployed cyber-physical systems.

Simultaneously, the rapid proliferation of large language models (LLMs) has introduced new categories of security risks. Recent surveys show that LLMs are susceptible to backdoor insertion, jailbreak prompting, covert model steering, and other manipulation techniques that undermine reliability and accountability [22]. These vulnerabilities point to

the need for unified approaches that consider adversarial robustness, privacy leakage, and model governance together rather than in isolation.

Collectively, these emerging threats illustrate that AI risk is multifaceted, spanning accuracy, robustness, fairness, privacy, and security dimensions. While existing governance frameworks identify many of these concerns, they provide limited quantitative mechanisms for evaluating them across heterogeneous regulatory regimes. This motivates the need for an integrative evaluation method such as the CAVe framework proposed in this study.

Taken together, the existing body of work on AI governance and risk management can be broadly categorized into three complementary streams. The first stream focuses on regulatory and governance frameworks, including the NIST AI Risk Management Framework, the EU AI Act, and ISO/IEC standards, which provide high-level principles, risk categories, and organizational requirements for trustworthy AI deployment [1,2,5,11]. While these frameworks establish important normative and procedural foundations, prior studies consistently note the absence of explicit mechanisms for quantitatively assessing technical compliance in a unified and comparable manner.

The second stream addresses AI assurance, auditability, and documentation practices, such as internal algorithmic audits, model documentation, and governance processes, which aim to improve transparency and accountability. However, these approaches typically emphasize process compliance and qualitative evidence, rather than producing reproducible, metric-based evaluation outcomes that can be compared across regulatory regimes [10].

The third stream consists of technical studies on model-level properties, including robustness, privacy, and fairness, which propose quantitative evaluation methods under specific threat models or application assumptions [22,23]. Although these works provide valuable technical insights, they are generally developed in isolation from regulatory frameworks and do not offer a systematic mechanism for integrating heterogeneous regulatory requirements into a single compliance assessment.

As a result, a critical research gap remains between regulatory intent and technical evaluation: existing studies either define high-level governance requirements without quantitative scoring, or provide quantitative metrics without cross-framework regulatory alignment. This gap motivates the need for a policy-aware, quantitative evaluation framework that can bridge heterogeneous AI governance regimes while remaining grounded in measurable technical indicators.

3. Proposed Framework

The proposed CAVe framework maps key controls from the NIST AI RMF, the EU AI Act, and ISO/IEC 23894 and 42001 to a common set of indicators and quantifies them to compute the CCI [1,2,5,6]. This enables consistent evaluation of compliance levels across heterogeneous AI governance frameworks.

To ensure consistency across the framework description, the main notation used throughout this section is summarized in Table 1.

Table 1. Notation used throughout the CAVe framework.

m_k	Raw metric value for indicator k
$\tilde{m}_k$	Normalized (safe) metric value in $[0, 1]$
$s(\tilde{m}_k)$	Threshold-based partial score
$\alpha_{i,f}$	Internal weight of requirement i for framework f
β_f	Cross-framework weight
$\lambda_{i,f}$	Evidence-based penalty factor
CCI	Cross-Compliance Index

3.1. Framework Structure

CAVe evaluates AI system compliance in three steps:

1. Measure common metrics (accuracy, robustness, privacy, fairness) for each framework.
2. Normalize each metric to produce comparable subscores.
3. Aggregate the weighted normalized scores to compute the overall CCI, applying thresholds and veto rules for safety-critical validation.

The four core indicators selected in this study are derived through a procedural analysis that identifies the intersection of control items commonly present across the three frameworks and measurable in a quantitative manner [1,2,5,6]. First, the detailed requirements of the NIST AI RMF, the EU AI Act, and ISO/IEC 23894 and 42001 were categorized into functional groups, and each group was evaluated to determine whether it is reproducible through model-level measurement and whether the terminology and objectives are aligned across the frameworks [10,11]. As a result, Accuracy, Robustness, Privacy, and Fairness were identified as the largest common subset suitable for consistent quantitative evaluation.

In this study, we operationalize these four core indicators—collectively designated as the Structured AI Indicators (SAI-4). These indicators function as the technical pillars of trustworthy AI, selected to capture the intersection of critical requirements across heterogeneous regulatory frameworks:

- Accuracy establishes the functional baseline of the AI system; however, we interpret it as a necessary but insufficient condition, acknowledging prior findings that accuracy alone fails to guarantee reliability within complex or adversarial deployment environments [24,25].
- Robustness is consequently integrated as a complementary dimension to assess the stability of model behavior under input perturbations and environmental uncertainties, thereby addressing the safety risks inherent in real-world operations [24,25].
- Privacy quantifies the system's resistance to information leakage in response to growing data protection mandates; reflecting the diverse threat landscape that ranges from membership inference to attribute inference attacks [23,26], the framework adopts a metric specifically evaluating susceptibility to re-identification.
- Fairness serves as a critical social safeguard to prevent discriminatory outcomes; while recognizing that fairness is an inherently multifaceted concept with varying normative definitions [27], CAVe operationalizes it as a measurable, group-level performance disparity to ensure that compliance assessment remains both reproducible and objective.

By rigorous operationalization of these metrics, the proposed framework effectively bridges the gap between abstract regulatory principles and concrete technical evaluation.

This study intentionally adopts this minimal and measurable indicator set (SAI-4) as a starting point. In contrast, higher-level concepts such as safety, transparency, and accountability were excluded from the initial version despite their high importance. Prior research has shown that these indicators are difficult to quantify consistently due to variations in standards and operational contexts, and are often implemented through procedural or documentation-based controls rather than direct technical metrics [9–11]. Accordingly, CAVe focuses on indicators that are both explicitly referenced across multiple regulatory frameworks and empirically measurable, while remaining extensible to these higher-level governance indicators in future iterations.

3.2. Metric Normalization and Subscore Calculation

To compare indicators with different units, all measurements are normalized to $\tilde{m}_k \in [0, 1]$:

$$\tilde{m}_k = \begin{cases} \frac{m_k - L_k}{U_k - L_k}, & \text{Higher is safer (e.g., accuracy),} \\ 1 - \frac{m_k - L_k}{U_k - L_k}, & \text{Lower is safer (e.g., error rate).} \end{cases}$$

Here, m_k is the raw value, and L_k, U_k denote the lower and upper bounds.

This normalization approach follows common practices in multi-criteria decision analysis and AI evaluation, enabling consistent aggregation of heterogeneous metrics [24,25].

Based on the normalized values, partial scores $s(\tilde{m}_k)$ are combined using logical aggregation operators (AND/OR) or weighted averaging (W-AVG) to compute the fulfillment value $a_{i,f}$ for each requirement i under framework f . These aggregation mechanisms are designed to reflect different regulatory semantics, such as mandatory constraints or compensatory trade-offs, as observed in existing AI governance and assurance practices [9,10].

Each raw metric m_k is measured in accordance with the evaluation principles defined in the Each indicator (Accuracy, Robustness, Privacy, and Fairness) is measured by considering both the primary and supporting indicators defined in the NIST AI RMF, the EU AI Act, and ISO/IEC 23894 [1,2,5,6]. To ensure consistent interpretation across frameworks, Table 2 summarizes the unified measurement indicators, their normalization directions, and their explicit or partial coverage in each framework.

Table 2. Integrated Measurement Definition and Framework Mapping.

Category	Indicator/Unit	Measurement Protocol and Data Source	Norm. Dir.	NIST	EU AI Act	ISO
<i>Technical</i>						
Accuracy	Error rate, F1, AUC	IID validation set performance, k -fold CV	↑	✓	△	✓
Robustness	OOD perf. drop (%), attack succ. rate (%)	Distribution-shift and adversarial tests (PGD/FGSM)	↓	✓	△	✓
Security	Vulnerability count, MTTR, incidents	Security scan report, incident response log		✓	△	✓
Privacy	PII leakage risk ratio (%)	Membership inference, encryption/decryption audit	mixed	✓	✓	△
Fairness	Max group gap (%), SPD/EO diff.	Groupwise performance comparison (balanced sample)	↓	✓	△	✓
Transparency	Log coverage (%), XAI score	Mandatory event log coverage, XAI quality check	↑	△	✓	✓
<i>Regulatory</i>						
Human Oversight	Intervention rate, bypass fail rate	Human-in-the-loop stop/retry testing	↑	△	✓	✓
Post-Market Monitoring	Alert sens./prec., report delay	Operational telemetry, PMM compliance rate	mixed	△	✓	✓
<i>Governance</i>						
Governance/Auditability	Doc. completeness, traceability (%)	Policy documentation, role trace logs, audits	↑	△	△	✓
Lifecycle Management	Re-eval. cycle, review rate	Stage-wise risk review, approval record (ISO 23894)	↑	△	△	✓

↑ Higher is safer; ↓ Lower is safer; mixed: context dependent. Symbols: ✓ = explicit coverage; △ = partial/implicit coverage in framework controls.

In this study, CAVe operationalizes four core indicators—Accuracy, Robustness, Privacy, and Fairness—as the primary inputs to the CCI, selected for their cross-framework commonality and stable quantitative measurability. As summarized in Table 2, additional technical, regulatory, and governance-related indicators (e.g., security, transparency, human oversight, and lifecycle management) are included to illustrate the broader measurement space defined across the three frameworks. These indicators are not directly aggregated into the CCI in the current implementation, but are treated as extensible dimensions that can be incorporated in future iterations as standardized quantitative metrics and regulatory practices mature.

The measured values are normalized and evaluated in a framework-aware manner, and the final compliance score is calibrated through weighted aggregation to reflect the relative importance of different regulatory perspectives. This design enables consistent cross-framework verification while mitigating bias toward any single standard. All raw measurements are normalized to the range $[0, 1]$ to ensure comparability across frameworks [24,25].

Each normalized indicator value $\tilde{m}_k$ is evaluated against a framework-specific threshold interval $(\theta_{k,i,f}^{\text{lo}}, \theta_{k,i,f}^{\text{hi}})$:

$$s(\tilde{m}_k) = \text{clip}\left(\frac{\tilde{m}_k - \theta_{k,i,f}^{\text{lo}}}{\theta_{k,i,f}^{\text{hi}} - \theta_{k,i,f}^{\text{lo}}}, 0, 1\right),$$

where values below $\theta_{k,i,f}^{\text{lo}}$ yield a score of 0, values above $\theta_{k,i,f}^{\text{hi}}$ yield a score of 1, and intermediate values are linearly interpolated. This thresholding mechanism enforces minimum regulatory requirements while preserving continuity and comparability of scores across heterogeneous frameworks [9,11].

3.3. Evidence-Based Penalty and Framework Scoring

To account for incomplete or unverifiable evidence, an evidence-based penalty factor $\lambda_{i,f} \in [0, 1]$ is applied to the fulfillment score:

$$a_{i,f} \leftarrow (1 - \lambda_{i,f}) a_{i,f}.$$

This penalty mechanism reflects established practices in AI assurance and auditability, where missing documentation or insufficient validation reduces confidence in compliance claims [10,28].

The framework-level compliance score is then computed as a weighted sum of penalized requirement scores:

$$C_f = \sum_i \alpha_{i,f} a_{i,f}, \quad \sum_i \alpha_{i,f} = 1,$$

where $\alpha_{i,f}$ represents the relative importance of requirement i within framework f . These weights enable the scoring process to reflect different regulatory priorities and compliance philosophies across frameworks, while maintaining a consistent mathematical structure for cross-framework comparison.

The threshold intervals, internal weights, cross-framework weights, and evidence-based deduction factors are derived through a normative procedure that reflects both the requirement levels and the distribution of control items across the three regulatory frameworks. Specifically, threshold ranges are defined by integrating the minimum performance requirements specified in the EU AI Act, the profile-based recommendations of the NIST AI RMF, and the risk acceptance criteria outlined in ISO/IEC 23894, thereby establishing lower and upper performance bounds that should be commonly satisfied across frameworks [1,5,6].

The internal weights $\alpha_{i,f}$ are determined based on the relative proportion and coverage of control domains within each framework, ensuring that the weights reflect both the density and the normative importance of the corresponding requirements, rather than being optimized against empirical performance outcomes [11].

The cross-framework weights β_f are assigned according to a domain-specific policy principle. In this study, highly regulated sectors such as healthcare and finance place greater emphasis on the EU framework, whereas technically oriented sectors such as manufacturing or defense prioritize the NIST framework. Accordingly, β_f functions not as a tunable optimization parameter but as a normative input that encodes domain characteristics and regulatory priorities, consistent with prior discussions on policy-driven AI risk management [9].

Furthermore, documentation completeness and evidence sufficiency criteria defined in ISO/IEC 42001 are incorporated such that missing, incomplete, or unverifiable evi-

dence results in partial deductions via the penalty factor $\lambda_{i,f}$, aligning the scoring process with established practices in AI governance, auditability, and management-system compliance [2,10].

3.4. CAVe Algorithm and Final CCI Computation

Algorithm 1 formalizes the CAVe scoring process by computing the CCI from the normalized indicators, threshold-based partial scores, evidence penalties, and framework weights. It expresses the stepwise computation flow implied by the formulations in Sections 3.2 and 3.3, showing how each intermediate component contributes to the final compliance score.

Inputs:

- Raw metric values m_k (normalized to $\tilde{m}_k$ inside the algorithm)
- Threshold parameters $(\theta^{lo}, \theta^{hi})$
- Internal requirement weights $\alpha_{i,f}$
- Cross-framework weights β_f
- Evidence-based penalty factors $\lambda_{i,f}$

Output:

- Final CCI
- Assigned grade (Pass/Conditional/Fail)

To complement the pseudocode in Algorithm 1, the overall computation pipeline is visualized in Figure 1, which shows how the normalization, scoring, penalty application, and grading steps are executed within the CAVe framework.

The final CCI formula is:

$$CCI_{\text{final}} = \sum_{f \in \{\text{NIST}, \text{EU}, \text{ISO}\}} \beta_f C_f, \quad \sum_f \beta_f = 1, \beta_f \geq 0.$$

Here, β_f represents the framework-level weighting—higher for EU in compliance-focused domains, whereas higher for NIST in technically oriented sectors.

The mathematical formulation of the CCI is proposed by the authors as an integrative evaluation mechanism. While individual components such as metric normalization follow standard practices, the unified integration of framework-specific thresholds, evidence-based penalty factors, and cross-framework weighting into a single computable index is an original contribution of this study. This formulation is specifically designed to enable policy-aware and reproducible compliance evaluation across heterogeneous AI governance frameworks.

Algorithm 1 CAVe Algorithm for CCI Calculation

Require: Raw metrics $\{m_k\}$, configs $\{(L_k, U_k, dir_k)\}$, frameworks $\mathcal{F} = \{\text{NIST}, \text{EU}, \text{ISO}\}$, weights $\{\alpha_{i,f}\}, \{\beta_f\}$

Ensure: CCI score and grade

- 1: **for** each metric k **do**
 - 2: Normalize $\tilde{m}_k$ according to dir_k and clip to $[0, 1]$
 - 3: **end for**
 - 4: **for** each requirement $r_{i,f}$ in framework f **do**
 - 5: Compute partial scores $s(\tilde{m}_k)$ for $k \in K_{i,f}$
 - 6: Aggregate via AND/OR/W-AVG $\rightarrow a_{i,f}$
 - 7: Apply evidence deduction $(1 - \lambda_{i,f})$
 - 8: **end for**
 - 9: $C_f = \sum_i \alpha_{i,f} a_{i,f}$; $CCI = \sum_f \beta_f C_f$
 - 10: Apply grading thresholds and veto rule
-

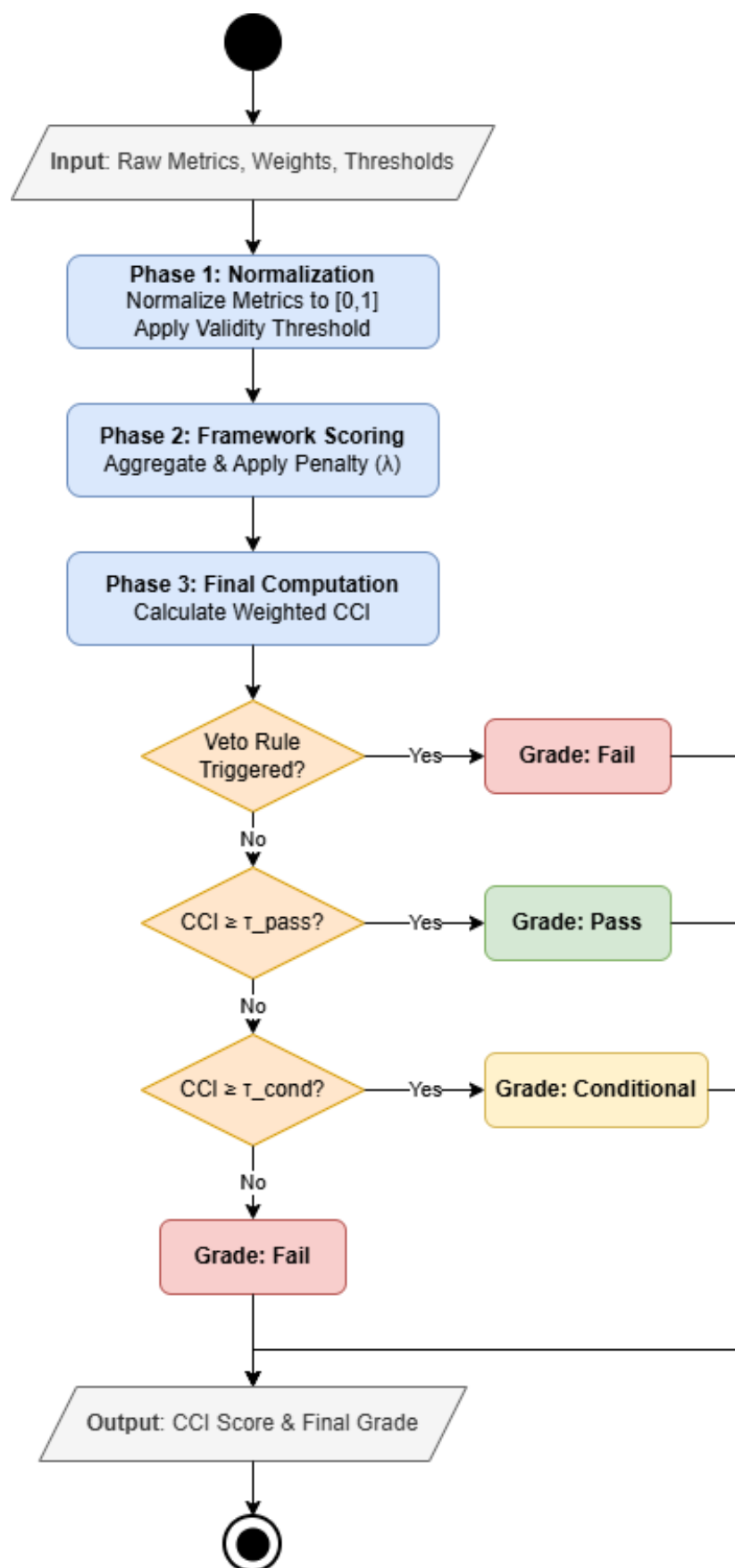


Figure 1. UML activity diagram illustrating the full computation workflow of the CAVe algorithm, including normalization, scoring, penalty application, and final grade determination.

Before computing the CCI, predefined veto conditions (e.g., major privacy breaches or legal non-compliance) are checked. If any are triggered, the evaluation is immediately labeled as Fail; otherwise, the grade is determined as follows:

$$\begin{aligned} \text{CCI} &\geq \tau_{\text{pass}} && \Rightarrow \text{Pass}, \\ \tau_{\text{cond}} &\leq \text{CCI} < \tau_{\text{pass}} && \Rightarrow \text{Conditional}, \\ \text{CCI} &< \tau_{\text{cond}} && \Rightarrow \text{Fail}. \end{aligned}$$

As an illustrative case, given $C_{\text{NIST}} = 0.72$, $C_{\text{EU}} = 0.65$, and $C_{\text{ISO}} = 0.81$ with equal weights ($\beta_f = \frac{1}{3}$):

$$\text{CCI} = 0.73,$$

which corresponds to a Pass grade under $\tau_{\text{pass}} = 0.7$ and $\tau_{\text{cond}} = 0.5$.

4. Validation

To verify the practical applicability of the proposed CAVe framework, we present two case studies. The first case deals with a common information security application, spam mail filtering, and the second focuses on a high-risk application domain, healthcare AI. Through these two cases, the differences in weight configuration and the application of the Veto Rule according to domain characteristics are examined.

In both scenarios, we applied a validity threshold range of $[0.6, 1.0]$ to the scoring function defined in Section 3.2. By setting the lower bound $\theta^{lo} = 0.6$ and the upper bound $\theta^{hi} = 1.0$, any performance metric falling below 0.6 is considered unacceptable and assigned a score of 0. This strict cutoff ensures that the CCI reflects only meaningful performance levels, filtering out sub-standard models while maintaining linearity for valid scores (≥ 0.6).

4.1. Materials and Methods

This subsection describes the datasets, model configurations, evaluation procedures, and licensing information used in the validation experiments. Two representative scenarios were evaluated: a low-risk text-classification model and a high-risk healthcare diagnostic model.

1. **Datasets:** The spam-classification experiment uses the UCI SMS Spam Collection dataset, which is released under the Creative Commons Attribution 4.0 (CC BY 4.0) license. The healthcare experiment uses the MIMIC-CXR dataset, which is distributed under the PhysioNet credentialed Data Use Agreement (DUA) and may be used only for approved research purposes. Both datasets were used in accordance with their respective licensing conditions.
2. **Model configuration:** In the spam scenario, a Multinomial Naïve Bayes classifier was trained using a Bag-of-Words representation extracted from the SMS corpus. In the healthcare scenario, a CheXNet-style convolutional neural network was applied to chest X-ray images from the MIMIC-CXR dataset.
3. **Evaluation procedure:** Each model was evaluated using the four indicators defined in the CAVe framework: accuracy, robustness, privacy, and fairness. Robustness was measured through input perturbation in the spam scenario and through distribution-shift evaluation in the healthcare scenario. Fairness was assessed based on group-level performance comparisons. Privacy risk was estimated using generalization-gap-based re-identification susceptibility and membership-inference tendencies.
4. **CCI computation:** All raw measurements were normalized to the range $[0, 1]$ following the procedure described in Section 3. Threshold scoring, evidence-based penalties, and framework-level weights were then applied to compute the final CCI for each experiment.

Dataset licensing and additional usage information have been added to the Data Availability section in accordance with the journal's author guidelines.

4.2. Spam Mail Filtering

In this validation scenario, we evaluated a spam classification system using the UCI SMS Spam Collection dataset [29]. The dataset was downloaded and preprocessed within the validation script, and a Multinomial Naive Bayes classifier was trained using a Bag-of-Words representation. A 70:30 train–test split produced 1672 test samples.

The model achieved an accuracy of

$$\text{Accuracy} = 0.9850.$$

To evaluate robustness against corrupted inputs, a character-level perturbation of 10% was injected into each test message. Using the same model and vectorizer, the perturbed test set achieved

$$\text{Accuracy}_{\text{noisy}} = 0.9797, \quad \text{Robustness} = \frac{\text{Accuracy}_{\text{noisy}}}{\text{Accuracy}} = 0.9945.$$

Fairness was measured by comparing the False Positive Rates (FPRs) of two groups based on message length: short messages (<50 characters) and long messages (≥ 50 characters). Let $\text{FPR}_{\text{short}}$ and FPR_{long} denote the FPRs of the two groups. The fairness indicator is defined as

$$\text{Fairness} = 1 - |\text{FPR}_{\text{short}} - \text{FPR}_{\text{long}}|,$$

and the measured value was

$$\text{Fairness} = 0.9908.$$

Privacy was evaluated using the generalization gap between training and testing accuracy, following empirical membership inference analyses in prior studies [30]. Let $\text{ACC}_{\text{train}}$ denote the training accuracy; then

$$\text{Privacy} = 1 - \max(0, \text{ACC}_{\text{train}} - \text{Accuracy}) = 0.9922.$$

Although the CAVe framework theoretically supports adjustable threshold intervals $(\theta^{lo}, \theta^{hi})$ for domain customization, this validation study applies a baseline configuration where $\theta^{lo} = 0$ and $\theta^{hi} = 1$. This approach adopts direct normalization to minimize subjective bias during the initial evaluation and ensures that the CCI reflects the raw performance metrics as objectively as possible. The measured metrics are summarized in Table 3.

Table 3. Measured metrics (SMS Spam Filtering).

Metric k	Raw Value m_k	Normalized $\hat{m}_k$
Accuracy	0.9850	0.9850
Robustness	0.9945	0.9945
Fairness	0.9908	0.9908
Privacy	0.9922	0.9922

Following the scoring method defined in Section 3, each framework score C_f was computed using uniform internal weights $\alpha_{i,f} = 0.25$. A 20% evidence penalty was applied to the EU Privacy metric in accordance with documentation sufficiency requirements. Using equal inter-framework weights $\beta_{\text{NIST}} = \beta_{\text{EU}} = \beta_{\text{ISO}} = 1/3$, the final CCI was computed as

$$\text{CCI} = 0.9741.$$

Under the strict acceptance threshold ($\tau_{pass} = 0.80$) established in this study, this score corresponds to a definitive Pass grade. This demonstrates that the CAVe Framework can perform complete, evidence-based compliance evaluation using real-world data, without relying on assumed or hypothetical risk values.

4.3. Healthcare AI

The healthcare domain is strongly influenced by EU regulations, and the protection of patient data and the safety of diagnostic processes are treated as the most critical requirements. Accordingly, this study assigns greater weight to the EU framework, setting the inter-framework weights to $(\beta_{NIST}, \beta_{EU}, \beta_{ISO}) = (0.25, 0.50, 0.25)$, as shown in Table 4.

Table 4. Inter-framework weights β_f (Healthcare AI).

Framework	Weight β_f
NIST	0.25
EU	0.50
ISO	0.25

This healthcare scenario assumes that a CheXNet model trained on the MIMIC–CXR dataset is used to diagnose chest X-ray images. Accuracy and Robustness were derived from the lesion-wise AUROC results presented in Rajpurkar et al. [31]. In that study, the Pneumonia AUROC of CheXNet (ours) was reported as 0.7680, and this value was used as the Accuracy metric:

$$\text{Accuracy} = 0.7680.$$

The variability of lesion-wise model performance was considered the Robustness metric. Among the reported AUROC scores, the value was highest for Emphysema at 0.9371 and lowest for Infiltration at 0.7345. Thus, the internal performance deviation was computed as 0.2026, and the Robustness metric was calculated as:

$$\text{Robustness} = 1 - (0.9371 - 0.7345) = 0.7974.$$

Privacy and Fairness were determined with reference to prior studies. Regarding Privacy, membership inference attack success rates for high-risk medical datasets typically range from 60–80% [30]. Conservatively reflecting this risk, we defined the re-identification likelihood as 0.25 (pure risk above random guessing). Thus, the Privacy score is defined as:

$$\text{Privacy} = 1 - 0.25 = 0.7500.$$

Regarding Fairness, Chen et al. [32] reported a maximum true positive rate (TPR) disparity of 9.3% across protected attributes in a MIMIC–CXR model. Thus, the Fairness score is defined as:

$$\text{Fairness} = 1 - 0.093 = 0.9070.$$

The measured metrics are summarized in Table 5. Consistent with the spam filtering scenario, direct normalization was applied.

Table 5. Measured metrics (Healthcare AI).

Metric k	Raw Value m_k	Normalized $\hat{m}_k$
Accuracy	0.7680	0.7680
Robustness	0.7974	0.7974
Fairness	0.9070	0.9070
Privacy	0.7500	0.7500

In healthcare AI, a Safety Veto Rule based on the diagnostic error rate ($1 - \text{Accuracy}$) acts as a critical safeguard. For example, if the safety threshold is set to 10%, a model with an error rate exceeding this limit is immediately rejected.

Following the updated scoring logic with a validity threshold range of $[0.6, 1.0]$, each framework score was computed using uniform internal weights ($\alpha = 0.25$). A 20% evidence penalty was applied to the EU Privacy metric. Based on the framework weights ($\beta_{\text{EU}} = 0.50$), the final CCI is calculated as:

$$\text{CCI} = 0.5046.$$

This score falls directly into the quantitative Fail range ($\text{CCI} < 0.60$), as the validity threshold strictly penalizes metrics closer to the lower bound (0.6). Moreover, the Safety Veto Rule was concurrently triggered because the diagnostic accuracy (76.80%) failed to meet the required safety threshold (90%). Consequently, the final decision is a definitive Fail. This outcome demonstrates that the CAVe framework effectively filters out high-risk models through a dual safeguard mechanism: rigorous quantitative scoring based on validity thresholds and absolute safety constraints via the Veto Rule.

Nevertheless, the proposed framework provides an objective assessment outcome that clearly reveals the quantitative levels of individual metrics and their regulatory alignment.

5. Discussion

This section analyzes how the CAVe framework responds to changes in (1) the measured indicator values, (2) the global validity threshold, and (3) the cross-framework weights. The objective is to demonstrate that the CCI is not merely a static aggregation score but a policy-sensitive evaluation mechanism that reflects both technical model performance and domain-specific regulatory priorities.

5.1. Experiment 1: Baseline Metric Sensitivity

We first examined how the CCI changes when individual metrics (accuracy, robustness, privacy, fairness) are increased while all other conditions remain fixed. As shown in Figure 2, the CCI grows linearly with increases in each indicator. This behavior is expected given the normalization and aggregation structure of CAVe, and it confirms that the framework maintains monotonicity and interpretability in its core scoring mechanism.

This baseline observation serves as a reference point: higher performance yields higher compliance, and no unexpected nonlinear effects appear when only raw metrics change. The more meaningful behaviors emerge when policy parameters—rather than the metrics themselves—are varied.

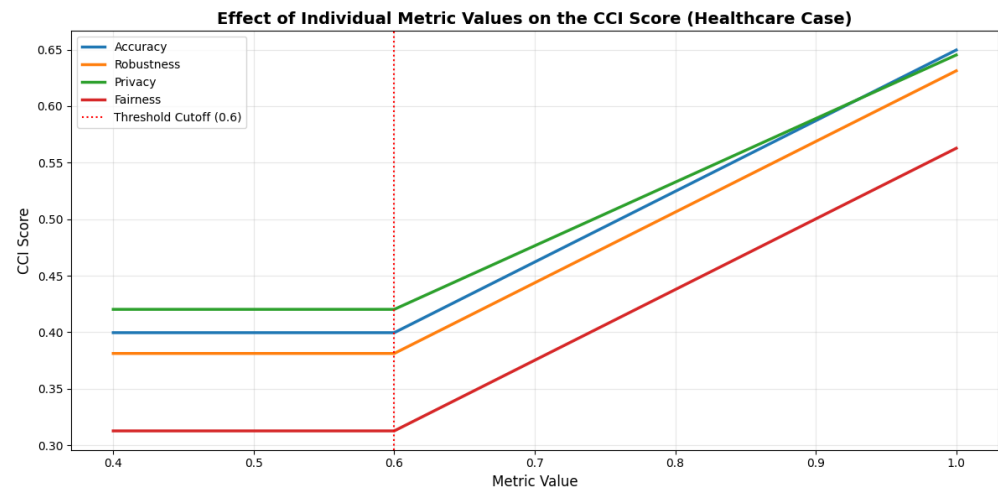


Figure 2. Effect of Individual Metric Values on the CCI Score (Healthcare Case).

5.2. Experiment 2: Threshold Sensitivity Analysis

A global lower validity threshold θ^{lo} was applied to all four indicators and gradually increased from 0.0 to 0.8. This threshold operates as a minimum acceptable performance criterion. When a metric value falls below θ^{lo} , its partial score is immediately reduced to zero, sharply reducing the final CCI.

Figure 3 illustrates this cutoff effect: once θ^{lo} surpasses a model's weakest indicator (e.g., Privacy = 0.75 in the healthcare case), the CCI exhibits a sudden drop. This behavior mirrors a “quality control” mechanism in which a tightening of global standards can instantly invalidate otherwise well-performing models.

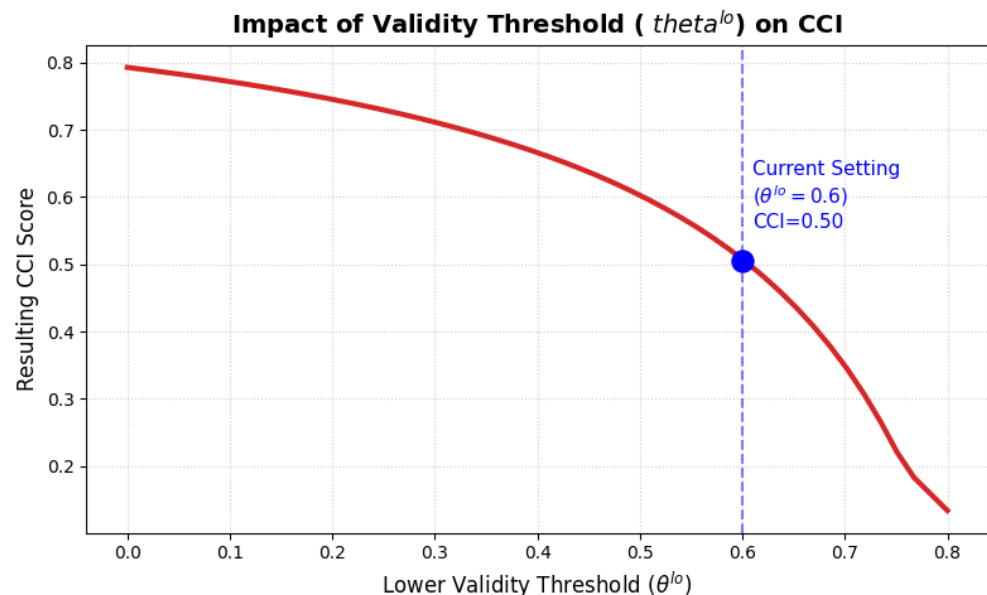


Figure 3. CCI sensitivity to the lower validity threshold θ^{lo} (current setting marked).

Conceptually, θ^{lo} functions as an absolute safety filter. It determines how strict the evaluation should be and ensures that every indicator meets at least the minimum acceptable level. This is particularly critical in high-risk domains such as healthcare, where failing a single requirement may be sufficient to reject a system.

5.3. Experiment 3: Weight Sensitivity Analysis

While thresholds represent absolute minimum safety requirements, framework weights (β_f) represent relative regulatory priorities. To analyze this effect, the EU framework weight β_{EU} was varied from 0.1 to 0.9 while ensuring that the weights sum to 1.0.

As shown in Figure 4, increasing β_{EU} produces a gradual decline in the resulting CCI. This behavior arises because the healthcare model's lowest indicator—Privacy (0.75)—is emphasized more strongly as the EU framework gains influence. Since the EU places significant regulatory weight on privacy protection, documentation sufficiency, and human oversight, raising β_{EU} amplifies the impact of the weakest dimension.

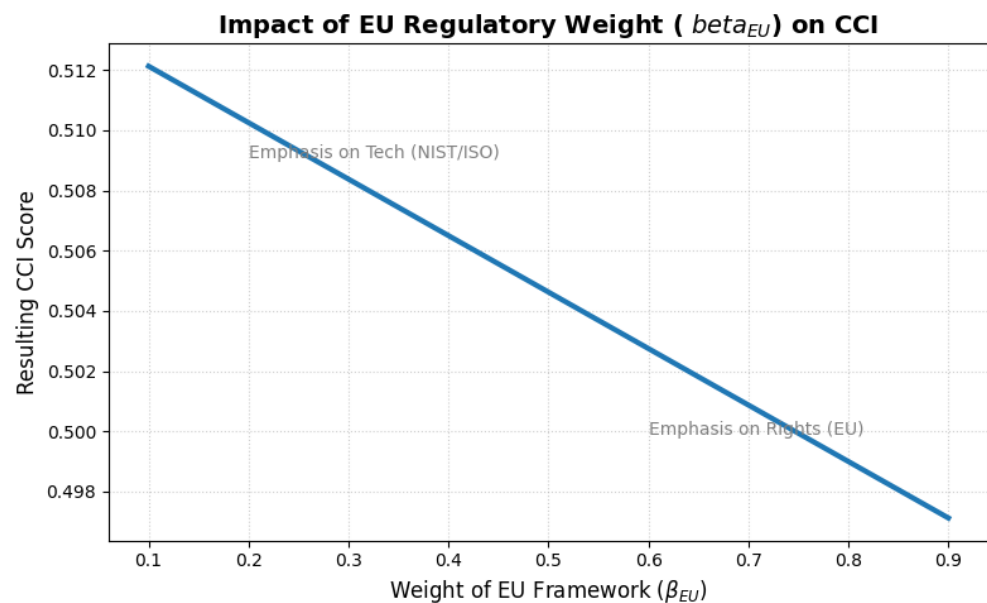


Figure 4. CCI variation as the EU framework weight β_{EU} increases.

This demonstrates that CAVe is capable of representing domain-specific regulatory stances. A model rated highly under a technical-performance-oriented regime (e.g., NIST-weighted) may score substantially lower under a rights-oriented regime (e.g., EU-weighted). Through β_f , CAVe provides a direct mechanism for encoding policy preferences into the evaluation itself.

5.4. Integrated Interpretation

This study is structured as an end-to-end technical and regulatory evaluation pipeline. Section 2 reviews existing AI governance frameworks, ontology-based approaches, and technical studies on AI security and privacy, and identifies the lack of a quantitative cross-framework compliance evaluation mechanism. Section 3 addresses this gap by introducing the CAVe framework, including indicator selection, metric normalization, threshold-based validation, and policy-aware aggregation into the CCI. Section 4 validates the proposed framework through controlled experiments that examine the effects of metric variation, regulatory thresholds, and framework-level weighting. Finally, this section discusses the implications of the proposed approach for reproducible, policy-aligned AI risk assessment across heterogeneous regulatory environments.

Together, the three experiments reveal that the CCI responds to both performance-driven and policy-driven factors:

- Metric changes produce predictable, linear improvements in the CCI, reflecting the technical performance of the AI model.
- Threshold changes produce discontinuous drops, acting as an absolute safety filter that enforces minimum regulatory criteria.

- Weight changes tune the evaluation outcome according to the regulatory philosophy of a specific jurisdiction or industry.

These results confirm that CAVe is not simply a multi-metric averaging scheme. Rather, it is a structured, policy-aware evaluation model that integrates both quantitative evidence and regulatory priorities.

In practice, this allows the same model to receive different CCI outcomes depending on the domain context: a model may pass under a performance-driven manufacturing-oriented assessment but fail under a privacy- and safety-critical healthcare assessment.

Therefore, the CAVe framework provides not only a reproducible and interpretable quantitative score but also a tunable mechanism for aligning compliance evaluations with legal, ethical, and operational requirements across heterogeneous regulatory environments.

This study adopts SAI-4 as a common set of indicators for quantitative cross-framework comparison. SAI-4 is selected as a structurally extensible foundation that captures the largest intersection of measurable requirements shared across the three international frameworks. Although higher-level indicators such as safety, transparency, human oversight, and accountability are critically important in real regulatory environments, existing policy and governance initiatives predominantly express these concepts in qualitative or procedural terms rather than through standardized quantitative metrics. Moreover, variations in evaluation methods and measurement protocols across jurisdictions further limit their direct integration into a unified scoring model. Accordingly, this study establishes a quantification structure centered on indicators that can be stably and reproducibly measured across all three frameworks, while maintaining architectural extensibility for incorporating additional indicators as regulatory quantification practices and domain-specific empirical evidence continue to mature.

From a policy perspective, the proposed CAVe framework can be interpreted as a technical bridge between value-oriented global AI principles and operational compliance assessment. International initiatives such as the OECD AI Principles [33] and the UNESCO Recommendation on the Ethics of AI [34] emphasize trustworthiness, fairness, and accountability, yet stop short of defining how these concepts should be measured quantitatively. By translating shared regulatory concerns into a structured set of measurable indicators, CAVe provides a pragmatic mechanism for aligning high-level ethical objectives with technical evaluation practices.

6. Conclusions

The proposed CAVe framework can quantitatively assess AI risk levels and regulatory compliance across different standards by mapping key controls from the NIST AI RMF, EU AI Act, and ISO/IEC 23894/42001 onto four core indicators—accuracy, robustness, privacy, and fairness. This enables consistent comparison of how AI systems meet various regulatory requirements.

Our scenario-based validation shows that the framework captures each regulation's distinct characteristics and reveals how the four indicators relate to one another. We see several directions for future work. First, we plan to expand the indicator set to include safety, transparency, and accountability. Second, we will optimize the weighting scheme and validate the extended CCI model for industry-specific contexts.

This work contributes a quantitative foundation for AI risk evaluation and creates a bridge between international AI governance standards. We believe it offers value both for academic research and for practitioners working in policy and industry settings. Future work will include applying the CAVe framework to real-world AI systems in collaboration with industry partners to verify its generalizability.

Author Contributions: Conceptualization, Writing—Original draft, Methodology, Software, visualization, Project administration, C.-H.M.; Writing—Review and editing, Validation, D.-G.L.; Supervision, J.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea Government (MSIT) (No. RS-2025-02304842, Development of Cloud-based Adaptive Security Architecture and Interoperability Integration API).

Data Availability Statement: The datasets used in this study are publicly available. The UCI SMS Spam Collection dataset is released under the Creative Commons Attribution 4.0 (CC BY 4.0) license and can be accessed at <https://archive.ics.uci.edu/ml/datasets/sms+spam+collection> (accessed on 8 January 2026). The MIMIC-CXR dataset is distributed under the PhysioNet credentialed Data Use Agreement (DUA) and can be accessed at <https://physionet.org/content/mimic-cxr/> (accessed on 8 January 2026). All datasets were used strictly in accordance with their respective licensing terms.

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

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
AIMS	Artificial Intelligence Management System
AI RMF	AI Risk Management Framework
AIRO	AI Risk Ontology
AUC	Area Under the Curve
AUROC	Area Under the Receiver Operating Characteristic
C2AIRA	Concrete and Connected AI Risk Assessment
CAVe	Cross-Assessment & Verification for Evaluation
CC BY	Creative Commons Attribution
CCI	Cross-Compliance Index
CV	Cross-Validation
DUA	Data Use Agreement
EO	Equal Opportunity
EU	European Union
F1	F1 Score
FGSM	Fast Gradient Sign Method
FPR	False Positive Rate
GPAI	General-Purpose Artificial Intelligence
IEC	International Electrotechnical Commission
IID	Independent and Identically Distributed
ISO	International Organization for Standardization
LLM	Large Language Model
MTTR	Mean Time To Recovery
NIST	National Institute of Standards and Technology
OOD	Out-of-Distribution
PGD	Projected Gradient Descent
PII	Personally Identifiable Information
PMM	Post-Market Monitoring
SAI	Structured AI Indicators
SMS	Short Message Service
SPD	Statistical Parity Difference
TPR	True Positive Rate
UCI	University of California, Irvine
UML	Unified Modeling Language
XAI	Explainable AI

References

1. European Parliament. *EU AI Act: First Regulation on Artificial Intelligence*; European Parliament: Strasbourg, France, 2023.
2. ISO/IEC 42001:2023; Information Technology—Artificial Intelligence—Management System. ISO/IEC: Geneva, Switzerland, 2023. Available online: <https://www.iso.org/standard/81230.html> (accessed on 8 January 2026).
3. Vassilev, A.; Oprea, A.; Fordyce, A.; Anderson, H. Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations. In *NIST Trustworthy and Responsible AI NIST AI 100-2e2023*; National Institute of Standards and Technology (NIST): Gaithersburg, MD, USA, 2024. [\[CrossRef\]](#)
4. Shrestha, S.; Banda, C.; Mishra, A.K.; Djebbar, F.; Puthal, D. Investigation of Cybersecurity Bottlenecks of AI Agents in Industrial Automation. *Computers* **2025**, *14*, 456. [\[CrossRef\]](#)
5. National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2023; p. 100-1. Available online: <https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf> (accessed on 8 January 2026).
6. ISO/IEC 23894:2023; Information Technology—Artificial Intelligence—Guidance on Risk Management. ISO/IEC: Geneva, Switzerland, 2023. Available online: <https://www.iso.org/standard/77304.html> (accessed on 8 January 2026).
7. Veale, M.; Zuiderveen Borgesius, F. Demystifying the Draft EU Artificial Intelligence Act. *arXiv* **2021**, arXiv:2107.03721.
8. Hacker, P.; Engel, A.; Mauer, M. Regulating ChatGPT and other Large Generative AI Models. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), Chicago, IL, USA, 12–15 June 2023. [\[CrossRef\]](#)
9. Brundage, M.; Avin, S.; Wang, J.; Belfield, H.; Krueger, G.; Hadfield, G.; Khlaaf, H.; Yang, J.; Toner, H.; Fong, R.; et al. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. *arXiv* **2020**, arXiv:2004.07213. [\[CrossRef\]](#)
10. Morley, J.; Floridi, L.; Kinsey, L.; Elhalal, B. An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. *Sci. Eng. Ethics* **2020**, *26*, 2141–2168. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Schiff, D.; Rakova, B.; Ayesh, A.; Fanti, A.; Lennon, M. Explaining the Principles to Practices Gap in AI. *IEEE Technol. Soc. Mag.* **2021**, *40*, 81–94. [\[CrossRef\]](#)
12. National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*; NIST Trustworthy and Responsible AI: Gaithersburg, MD, USA, 2024.
13. Barrett, A.M.; Hendrycks, D.; Newman, J.; Nonnecke, B. Actionable guidance for high-consequence AI risk management: Towards standards addressing AI catastrophic risks. *arXiv* **2022**, arXiv:2206.08966.
14. Smith, G.; Stanley, K.D.; Marcinek, K.; Cormarie, P.; Gunashekar, S. *General-Purpose Artificial Intelligence (GPAI) Models and GPAI Models with Systemic Risk: Classification and Requirements for Providers*; RAND: Arlington, VA, USA, 2024.
15. Simonetta, A.; Paoletti, M.C. *ISO/IEC Standards and Design of an Artificial Intelligence System*; CEUR: Aachen, Germany, 2024.
16. Boza, P.; Evgeniou, T. Implementing AI Principles: Frameworks, Processes, and Tools. *INSEAD Work. Pap.* **2021**. [\[CrossRef\]](#)
17. Golpayegani, D.; Harshvardhan, J.; Pandit, D.L. AIRO: An Ontology for Representing AI Risks Based on the Proposed EU AI Act and ISO Risk Management Standards. *Semantic Web* **2022**, *55*, 51–65. [\[CrossRef\]](#)
18. Xia, B.; Lu, Q.; Perera, H.; Zhu, L.; Xing, Z.; Liu, Y.; Whittle, J. Towards Concrete and Connected AI Risk Assessment (C²AIRA): A Systematic Mapping Study. *arXiv* **2023**, arXiv:2301.11616.
19. Karras, D.A. On Modelling a Reliable Framework for Responsible and Ethical AI in Digitalization and Automation: Advancements and Challenges. *WSEAS Trans. Financ. Eng.* **2025**, *3*, 333–350. [\[CrossRef\]](#)
20. Cui, Q.; You, X.; Wei, N.; Nan, G.; Zhang, X.; Zhang, J.; Lyu, X.; Ai, M.; Tao, X.; Feng, Z.; et al. Overview of AI and communication for 6G network: Fundamentals, challenges, and future research opportunities. *Sci. China Inf. Sci.* **2025**, *68*, 171301. [\[CrossRef\]](#)
21. Yuan, S.; Xu, G.; Li, H.; Zhang, R.; Qian, X.; Jiang, W.; Cao, H.; Zhao, Q. FIGhost: Fluorescent Ink-based Stealthy and Flexible Backdoor Attacks on Physical Traffic Sign Recognition. *arXiv* **2025**, arXiv:2505.12045. [\[CrossRef\]](#)
22. Zhou, Y.; Ni, T.; Lee, W.B.; Zhao, Q. A Survey on Backdoor Threats in Large Language Models (LLMs): Attacks, Defenses, and Evaluations. *arXiv* **2025**, arXiv:2502.05224. [\[CrossRef\]](#)
23. Choquette-Choo, C.A.; Tramer, F.; Carlini, N.; Papernot, N. Label-Only Membership Inference Attacks. *arXiv* **2021**, arXiv:2007.14321. [\[CrossRef\]](#)
24. Upreti, R.; Pedro, G.; Lind, A.E.; Yazidi, A. Security and privacy in large language models: A survey. *Int. J. Inf. Secur.* **2024**, *23*, 2287–2314. [\[CrossRef\]](#)
25. Nastoska, A.; Jancheska, B.; Rizinski, M.; Trajanov, D. Evaluating Trustworthiness in AI: Risks, Metrics, and Applications Across Industries. *Electronics* **2025**, *14*, 2717. [\[CrossRef\]](#)
26. Hu, H.; Salic, Z.; Sun, L.; Dobbie, G.; Yu, P.S.; Zhang, X. Membership Inference Attacks on Machine Learning: A Survey. *arXiv* **2022**, arXiv:2103.07853. [\[CrossRef\]](#)
27. Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; Galstyan, A. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* **2021**, *54*, 115. [\[CrossRef\]](#)

28. Raji, I.D.; Smart, A.; White, R.N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; Barnes, P. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT '20), Barcelona, Spain, 27–30 January 2020. [CrossRef]
29. Almeida, T.A.; Hidalgo, J.M.G.; Yamakami, A. Contributions to the study of SMS spam filtering: New collection and results. In Proceedings of the 11th ACM Symposium on Document Engineering, Mountain View, CA, USA, 19–22 September 2011; pp. 259–262.
30. Shokri, R.; Stronati, M.; Song, C.; Shmatikov, V. Membership Inference Attacks against Machine Learning Models. *arXiv* **2017**, arXiv:1610.05820. [CrossRef]
31. Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; Shpanskaya, K.; et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv* **2017**, arXiv:1711.05225.
32. Chen, H.; Alfred, M.; Brown, A.D.; Atinga, A.; Cohen, E. Intersection of Performance, Interpretability, and Fairness in Neural Prototype Tree for Chest X-Ray Pathology Detection: Algorithm Development and Validation Study. *JMIR Form. Res.* **2024**, *8*, e59045. [CrossRef] [PubMed]
33. OECD. OECD Principles on Artificial Intelligence. 2019. Available online: <https://oecd.ai/en/ai-principles> (accessed on 8 January 2026).
34. UNESCO. Recommendation on the Ethics of Artificial Intelligence. 2021. Available online: <https://unesdoc.unesco.org/ark:/48223/pf0000380455> (accessed on 8 January 2026).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.