The Symbolic Blindspot: Measuring the Gap Between AI Governance Frameworks and Real-World Failure in Rule-Based Systems

Habiba Effat

Abstract:

Paradigm-Agnostic governance frameworks promise oversight for AI in a “one-size-fits-all” approach–but AI is not a monolith, and treating it like one fails to address a very critical elephant in the room. Symbolic AI–a method that saw intelligence as the application of explicit rules and logic to the manipulation of human-readable symbols–was long dismissed as a mere stepping stone for its flashier, probabilistic successor–Neural AI. While the attention of researchers, policymakers, and industries has shifted toward the neural paradigm, Symbolic systems remain widely deployed–particularly in safety-critical areas such as healthcare, where compliance is non-negotiable. Yet, major AI governance frameworks (e.g., the EU AI Act, NIST AI RMF) are written in paradigm-agnostic language or focus primarily on high-risk neural AI, leaving the governance of modern, sophisticated symbolic systems a critically underexplored area.

This paper tests whether the technical requirements of these frameworks produce meaningful assessments when applied to rule-based systems through a two-layer evaluation. First, governance requirements drawn from the EU AI Act (Articles 9–15, 72) and NIST AI RMF functions are operationalized into a three-point applicability score, measuring whether each requirement is satisfied natively by symbolic vs neural architectures. Second, these requirements are cross-referenced against documented, real-world failure modes of 103 rule-based clinical decision support malfunctions---to test whether theoretical applicability predicts operational reliability.

In our first layer of evaluation, Symbolic systems scored substantially higher compared to their neural counterparts—both scoring an average mean of 1.55 and 1.0, respectively. A result driven primarily by native traceability and oversight mechanisms. However, this advantage didn’t hold operationally: over a quarter of malfunctions fell into categories that weren't natively addressed by governance requirements—most notably conceptualization, which was the highest overall reported malfunction cause included in the study (20.3%). These findings reveal a crucial gap between theoretical governance applicability and real-world reliability for rule-based symbolic systems.

Keywords: AI governance; Symbolic AI; Trustworthy Artificial Intelligence; AI Assurance; EU AI Act; NIST AI RMF

Introduction:

AI governance has entered a new era, rapidly transforming in the span of 3 years from mere aspirations to dynamic multi-framework regulations. The main players in the AI governance landscape include the European Union's AI Act, the world's first comprehensive AI regulation, and the U.S. National Institute of Standards and Technology's AI Risk Management Framework (NIST AI RMF), which impose binding and quasi-binding regulations on AI practitioners (Tanveer 2026). While these frameworks diverge between judicial contexts and granularity, they share a "one-size-fits-all" approach when it comes to different AI paradigms. These frameworks are written, almost universally, in paradigm-agnostic language (Ali and Dornaika 2025), where "AI" is treated as a monolith, and risk is defined by critical application domains (education, employment, etc.) rather than by the underlying computational architecture. (Regulation (EU) 2024/1689, as amended by the Digital Omnibus on AI, Regulation (EU) 2026/1744)

Symbolic AI, a rule-based approach rooted in explicit logic, was a crucial stepping stone in the early days of AI development. However, its limitations grew evident when encountering uncertainty. This pushed research toward probabilistic neural systems that learn patterns from data and make predictions without explicit programming logic—marking a clear paradigm shift (Kelvin D. O et al. 2025). Yet, the drawbacks of the symbolic paradigm are also its main selling points. Modern Symbolic systems remain deployed today, particularly in safety-critical domains (e.g., healthcare), where deterministic, auditable behaviour is preferred precisely because it is *not* probabilistic. By contrast, most modern, complex Neural systems today operate essentially as a "black box", meaning it is incredibly challenging to trace back the origins of *why* a decision was made or how a specific input relates to an output. This quality of vague uncertainty makes deployment of these systems extremely risky in domains like healthcare and autonomous driving where compliance is non-negotiable (Kelvin D. O et al. 2025). The distinction followed in this paper is drawn from the dual paradigm framework introduced by Ali and Dornaika (2025), which characterizes AI systems into two distinct paradigms: The symbolic/classical vs the neural/generative. Under this framework, these two paradigm approaches differ not only in performance and use cases but in what "compliance" and "effective governance" could even entail—governance of symbolic systems includes verifying their logical structures, while for neural systems the focus is on reviewing training data, prompts, and stochastic behaviour. (Ali and Dornaika 2025)

Currently, future research directions strongly point to hybridization, in which the competing paradigms are integrated to form hybrid neuro-symbolic systems that inherit the high explainability and logical consistency of symbolic systems and the adaptability of their neural counterparts. But they also inherit each system's individual fallbacks and their respective governance challenges. This further highlights the necessity of an effective governance framework that accounts for both paradigms and their unique challenges. (Ali and Dornaika 2025) The current horizontal, paradigm-agnostic approach, which generalizes obligations across an umbrella of AI systems and application domains, ultimately remains too open-textured to be operationalized at scale. (Tamponi et al. 2026)

This leaves us with two questions: How do academic and policy conversations overwhelmingly center the neural paradigm—Large Language Models, Image recognition, etc.--- leaving behind

a huge gap, a Symbolic blind spot? And can these paradigm-agnostic regulation frameworks effectively regulate Symbolic rule-based systems?

Materials and Methods:

This paper follows a two-layer evaluation: First, Governance requirements are operationalized into a scoreable criterion to measure whether they maintain applicability irrespective of computational architecture—Neural vs Symbolic. Second, an operational layer is applied to measure whether that score holds against real-world documented failure modes.

Frameworks: Two governance frameworks were selected for a cross-framework analysis. Articles 9-15 and 72 of the EU AI Act (Regulation (EU) 2024/1689, as amended by the Digital Omnibus on AI, Regulation (EU) 2026/1744) and the 4 functions of the U.S. NIST AI Risk Management Framework (NIST AI 100-1), as mapped out in *Table 1*. These two frameworks represent the two main approaches when it comes to AI governance—binding regulations vs voluntary standards. Additionally, both are written intentionally using mostly paradigm-agnostic language or focus on high-risk Neural AI (Ali and Dornaika 2025), making them appropriate case studies to address our primary research question.

Challenge	EU AI Act	NIST AI RMF
Risk Management	Art.9	Manage 1.2,1.3, 2.2 Measure 3.1
Data governance	Art.10	Measure 2.11
Technical documentation & Traceability	Art.11	Map 1.1, 2.3 Measure 2.1
Logging/ Record Keeping	Art.12	Measure 2.4
Transparency & Explainability	Art.13	Measure 2.8, 2.9
Human oversight	Art.14	Map 3.5 Govern 3.2
Accuracy & Robustness	Art.15	Measure 2.5. 2.6
Security & Resilience (Cybersecurity)	Art.15	Measure 2.7
Post-Market Monitoring	Art. 72	Manage 4.1

Table 1
Cross-framework mapping of AI governance challenges, adapted from (Ugarte et al. 2026)

Layer 1: Each governance challenge from Table 1 was scored against the two architectures on a 3-point scale, accompanied by a 1-2 sentence rationale explaining the assigned score. :
0 = Not applicable/system architecture cannot natively satisfy this without fundamental redesign.
1 = Applicable with reframing/external tooling.
2 = Directly applicable/the architecture's native properties satisfy the requirement.

The mean score and differences of means for both systems were calculated to produce a numerical score that compares the applicability of governance for each of the two paradigms.

Layer 2: A simple numerical score only answers one part of the puzzle: it describes theoretical applicability, not reliability in real-world settings. To evaluate whether Layer 1 holds operationality, each requirement was cross-referenced against documented failure modes of symbolic systems in literature.

To avoid domain and applicability differences, The Governance challenges were mapped against 103 documented malfunctions of rule-based Clinical Decision Support (CDS) systems deployed in healthcare, categorized under the taxonomy developed by Wright et al. (2018) and extended by Thayer et al. (2024). Full definitions and descriptions for each cause category are provided in Appendix Table 1A.

The general “software error” (n=27) umbrella category was excluded from our evaluation due to its ambiguity and unspecificity in the original study, precluding a definitive classification. Leaving us with 103 Malfunctions from the original 130 identified by Thayer et al. (2024).

Each Malfunction cause category was classified as mapped (a corresponding governance requirement exists), unmapped (no requirement addresses this failure mode, a governance gap), or partial (a corresponding governance requirement exists for some but not all failure modes under this category, and a definitive answer is contingent on facts not specified in the source material).

Results:

Layer 1:

Table 2 represents applicability scores for nine governance challenges drawn from the EU AI Act and NIST AI RMF, scored for neural and symbolic architectures on a 3-point scale (0 = not applicable, 1 = applicable with reframing/external tooling, 2 = directly applicable).

Across all nine governance challenges, symbolic systems scored a 1.55 mean applicability score, compared to 1.0 for neural systems—with a mean gap of 0.55 in favor of symbolic architectures. The scores were not evenly distributed: Symbolic systems scored a 2 (directly applicable) on five of the nine challenges (55.5%)---Risk management, Technical Documentation & Traceability, Logging & Record Keeping, Human Oversight; Conversely, Neural systems scored a consistent 1 (applicable with reframing/external tooling) across all nine challenges.

Four challenges—Data Governance, Accuracy & Robustness, Security & Resilience (Cybersecurity), Post-Market Monitoring----together produced an identical score of 1 for both paradigms, representing a *joint gap.* This gap accounts for 44.4% of total governance challenges that are not solved by architecture choice alone. Explicit rule logic and deterministic behaviour directly satisfy documentation, traceability, and logging; But neither paradigm possesses inherent mechanisms to ensure that data is free from bias, performance holds under adversarial or novel conditions, or that correctness persists post-deployment.

challenge	eu_art	nist_func	symbolic_score	neural_score	Rationale
Risk Management	Art. 9	Manage 1.2, 1.3, 2.2; Measure 3.1	2	1	Symbolic: allows systematic analysis of deterministic behaviour. Neural: requires ongoing monitoring of probabilistic behaviour
Data Governance	Art. 10	Measure 2.11	1	1	Symbolic: bias emerges from explicit hand-coded rules or knowledge bases, easier to identify but hard to root out if

					foundational. Neural: arises from biases in training data and is amplified stochastically (Ali and Dornaika 2025)
Technical Documentation & Traceability	Art. 11	Map 1.1, 2.3; Measure 2.1	2	1	Symbolic: can be documented directly. Neural: requires extensive documentation of the model development cycle
Logging & Record Keeping	Art. 12	Measure 2.4; Manage 4.1	2	1	Symbolic: direct recording of rule execution. Neural: continued logging of system input, output, model versions, etc., requiring specific logging infrastructure
Transparency & Explainability	Art. 13	Measure 2.8, 2.9	2	1	Symbolic: inherently high by design; Neural: inherently low “black boxes” without post-hoc explainable AI techniques and decision logs (Kelvin D. O et al. 2025)
Human Oversight	Art. 14	Map 3.5; Govern 3.2	2	1	Symbolic: typically designed as explicit veto points or permission gates within a deterministic loop. Neural: more

					"fuzzy", require additional monitoring mechanisms (Ali and Dornaika 2025)
Accuracy & Robustness	Art. 15	Measure 2.5, 2.6	1	1	Symbolic: "brittle" struggle with messy real world/ missing data and uncertainty. Neural: vulnerable to adversarial attacks and distribution shifts (Tamponi et al. 2026)
Security & Resilience (Cybersecurity)	Art. 15	Measure 2.7	1	1	Vulnerabilities include: Symbolic: logic bombs & exploiting algorithmic flaws; Neural: prompt injection & training data poisoning (Ali and Dornaika 2025)
Post-Market Monitoring	Art.72	Manage 4.1	1	1	Symbolic rules require manual re-verification across environments, and Neural systems require equivalent external monitoring to catch post-deployment behavioral change.

Table 2
Applicability scores for nine governance challenges drawn from the EU AI Act and NIST AI RMF, scored for neural and symbolic architectures.

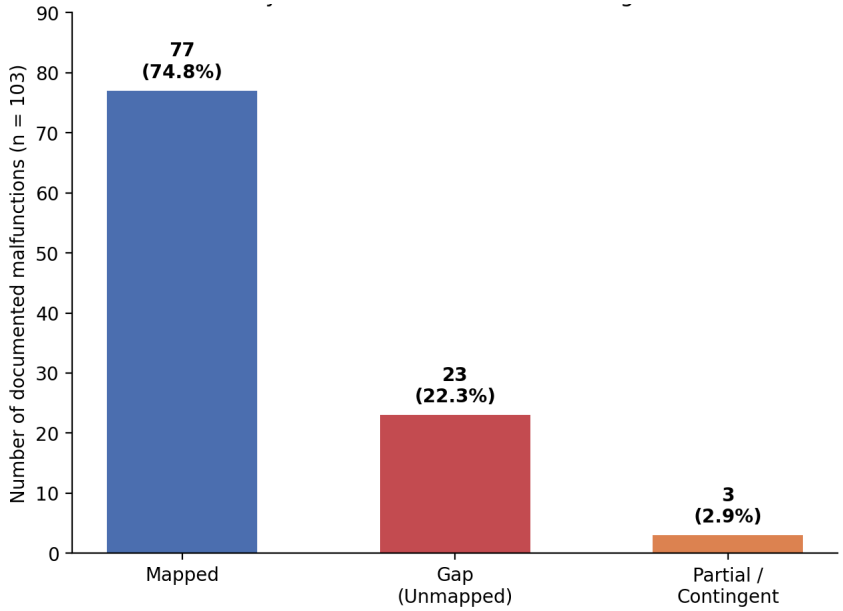
These applicability scores only measure which governance challenges each architecture can theoretically satisfy, but nothing about how this theoretical applicability holds in practice. The following section (Layer 2) tests this directly by mapping these same challenges against real-world documented failure modes in literature.

Layer 2:

Table 3 shows The Governance challenges mapped against 103 documented malfunctions of rule-based Clinical Decision Support (CDS) systems categorized under the taxonomy developed by Wright et al. (2018) and extended by Thayer et al. (2024). This breakdown is summarized in Figure 1.

Two cause categories—conceptualization (n=21) and Hardware error* (n=2)---were not directly mappable to the nine governance challenges identified (22.3%), and a third, organizational policy (n=3), showed only partial coverage contingent on facts not specified in the source material (2.9%). Together, these three categories account for 26 out of 103 of the malfunctions (25.2%), with over 80% belonging to the conceptualization category alone (the largest cause of total malfunctions overall).

The remaining mappable malfunctions greatly skewed towards one specific challenge—Accuracy & Robustness—which alone covered 62 out of 77 (80.5%) of mapped causes.



Mapping breakdown of the 103 documented rule-based CDS malfunctions (Thayer et al., 2024) as covered in Table 3

Hardware error was classified as unmapped; Article 15's only concrete elaborations concern algorithmic bias feedback loops and adversarial attack resilience, with no reference to physical infrastructure. *

Root cause of the malfunction	Maps to a governance challenge?	Specific challenge
Defect in EHR software (n=10)	Yes	Accuracy & Robustness
Conceptualization (n=21)	No	-----
Build error (n=18)	Yes	Accuracy & Robustness
New code, concept or term introduced, but rule not updated (n=15)	Yes	Accuracy & Robustness
Networking (n=6)	Yes	Accuracy & Robustness
Component failure (n=4)	Yes	Accuracy & Robustness
Information exchange (n=3)	Yes	Accuracy & Robustness
Environment migration (n=7)	Yes	Post-market monitoring
New value (n=6)	Yes	Accuracy & Robustness
Organizational policy (n=3)	Partial	Nearest adjacent (Art. 26), but applicability would depend on what the provider's instructions for use actually specify, which is unknown from the case description.
Inadvertent enabling/disabling	Yes	Human Oversight

(n=3)		
Alert text mismatch (n=3)	Yes	Transparency & Explainability
Hardware error (n=2)	No	---
Unaware of component reuse (n=1)	Yes	Technical Documentation & Traceability
Data source (n=1)	Yes	Data governance

Table 3
The Governance challenges mapped against 103 documented malfunctions of rule-based Clinical Decision Support (CDS) systems categorized under the taxonomy developed by Wright et al. (2018) and extended by Thayer et al. (2024).

Discussion:

Together, the two layers reveal a stark divergence between what compliance entails on paper and what that same compliance looks like in practice. Layer 1 found that Symbolic systems scored substantially higher than their Neural/statistical counterparts—scoring 1.55 and 1.0, respectively—across the governance challenges outlined in Table 2. This result is driven primarily by the native architecture of symbolic systems that supports explainability and human-readable logic. On the surface, this result suggests that coverage is not only accounted for but actually more rewarding for rule-based architectures.

Our second layer of analysis flips the script considerably. Of the 103 documented causes of malfunction, three of which—conceptualization errors, hardware errors, and organizational policy—were not fully covered by the governance requirements. Together, these three causes accounted for over a quarter of cases and included the most represented error cause overall: conceptualization errors (n=21), which accounted for 20.3% of the total 103 errors included in this study. The same Symbolic systems that scored promisingly in the first layer of analysis ultimately failed to account for the largest cause of malfunction: A misrepresentation of clinical concepts that, while perfectly traceable, remains flawed.

The remaining mappable malfunctions significantly skewed toward the Robustness & Accuracy governance challenge (represented by Art.15 and measures 2.5 & 2.6 of the EU AI Act and the NIST AI RMF, respectively), accounting for over 80% of total mapped malfunctions. A pattern directly attributable to the intentional vagueness observed as a direct result of the horizontal “one-size-fits-all” approach in governance: Article 15 requires an "appropriate level" of accuracy, robustness, and cybersecurity and resilience to be "as resilient as possible" to errors and faults, without a formal definition of robustness or operational guidance for demonstrating compliance in practice. This approach ultimately leaves Article 15 broad enough to plausibly capture nearly any performance-related failure and absorb a disproportionate share of any mapping exercise, leaving too much space for interpretation that could cause superficial compliance. (Tamponi et al. 2026)

Limitations: This study has several limitations that constrain the generalizability of its findings. First, all systems studied in Layer 2 operate within a single domain and application context—clinical decision support in healthcare. Whether the specific pattern of governance gaps identified here (unmapped areas and concentration of conceptualization-related malfunctions) generalizes to symbolic systems in other high-risk domains, such as credit scoring or criminal justice risk assessment, remains untested. Second, our analysis was conducted on a limited number of case studies: 130 malfunctions drawn from 28 published studies represent a small, non-random group of CDS malfunctions that happen to be reported. Malfunctions that were never published or simply fell out of the scope of the reviewed literature by Thayer et al., 2024 might not necessarily capture the same distribution of failure cases included in this study.

Conclusions:

This paper set out to test whether intentionally paradigm-agnostic governance requirements provide meaningful assessments when applied to symbolic rule-based systems or whether it was simply operating against a checklist built for another paradigm. The answer, based on our extensive two-layer analysis, is not a simple yes or no. At the level of theoretical applicability, symbolic systems scored more favorably relative to neural systems against the same governance challenges. Their explicit rule traceability, human-readable logic, and oversight mechanisms map clearly for requirements like explainability, logging, and human oversight in a way that native neural architectures fail to satisfy without external tooling.

Following our second layer of analysis, this specific advantage was found not to hold against real-world documented malfunctions. Over a quarter of the 103 documented Clinical Decision Support (CDS) system malfunctions didn't perfectly fit against any of the nine identified governance challenges—Including the most prevalent cause of the documented malfunctions (conceptualization errors).

This divergence between theoretical applicability and real-world reliability is this paper's main contribution. This gap is inherently structural rather than incidental: a horizontal framework broad enough to plausibly apply to any AI system risks providing little operational guidance for the paradigm-specific failure modes any single system actually exhibits. The EU AI Act recognises the need to account for context; however, current standardisation efforts largely focus on horizontal coverage and offer limited support for operationalising this requirement across contexts, including core architectural differences (Tamponi et al. 2026).

These findings carry important implications for policymakers, practitioners and standardisation bodies. Current regulatory standards—underdeveloped for symbolic systems—should be extended to address architecture-specific differences and challenges, rather than the widely adopted "one-size-fits-all" approach. Future work should extend this dual-layer framework across different high-risk application domains—including areas like credit scoring and criminal justice assessment—to measure the applicability of the specific gap identified here.

References

[1] Ali, Mohamad Abou, and Fadi Dornaika (arXiv.Org). 2025. “Agentic AI: A Comprehensive Survey of Architectures, Applications, and Future Directions.” October 29. https://doi.org/10.1007/s10462-025-11422-4.
[2] Kelvin D. O, Adjogri, Michael Uche Ikpade, and Sylvia Onataghogho Oyovwe. 2025. “From Symbolic AI to Generative Models: Tracing the Shifting Paradigms of Artificial Intelligence.” *Journal of Institutional Research, Big Data Analytics and Innovation* 1 (3): 147–60.
[3] Tamponi, Roberta, Carina Prunkl, Thomas Bäck, and Anna V. Kononova. 2026. “Lost in Vagueness: Towards Context-Sensitive Standards for Robustness Assessment under the EU AI Act.” arXiv:2511.15620. Preprint, arXiv, July 5. https://doi.org/10.48550/arXiv.2511.15620.
[4] Tanveer, Rizwan. 2026. “AI Governance Frameworks: ISO/IEC 42001, NIST AI RMF, and the EU AI Act.” SSRN Scholarly Paper No. 6740438. Social Science Research Network, May 9. https://doi.org/10.2139/ssrn.6740438.
[5] European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). *Official Journal of the European Union* L 2024/1689, July 12, 2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.
[6] National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1
[7] Ugarte, Rodrigo Cilla, Miguel Ángel Patricio Guisado, Antonio Berlanga de Jesús, and José Manuel Molina López. “Making AI Compliance Evidence Machine-Readable.” arXiv:2604.13767. Preprint, arXiv, April 15, 2026. https://doi.org/10.48550/arXiv.2604.13767.
[8] Thayer, Jeritt G., Amy Franklin, Jeffrey M. Miller, Robert W. Grundmeier, Deevakar Rogith, and Adam Wright. “A Scoping Review of Rule-Based Clinical Decision Support Malfunctions.” *Journal of the American Medical Informatics Association : JAMIA* 31, no. 10 (2024): 2405–13. https://doi.org/10.1093/jamia/ocae187.
[9] Adam Wright, Angela Ai, Joan Ash, Jane F Wiesen, Thu-Trang T Hickman, Skye Aaron, Dustin McEvoy, Shane Borkowsky, Pavithra I Dissanayake, Peter Embi, William Galanter, Jeremy Harper, Steve Z Kassakian, Rachel Ramoni, Richard Schreiber, Anwar Sirajuddin, David W Bates, Dean F Sittig, Clinical decision support alert malfunctions: analysis and empirically derived taxonomy, Journal of the American Medical Informatics Association, Volume 25, Issue 5, May 2018, Pages 496–506, https://doi.org/10.1093/jamia/ocx106

Appendix:

Root cause of the malfunction	desc.
Defect in EHR software	There is a programming error in the source code of the EHR that causes the EHR to function other than as designed or documented, and this error causes CDS to malfunction.
Conceptualization	There was an error in the clinical conceptualization of the rule, which, when implemented, led to a malfunction. Occurs when a rule is not designed correctly (eg, an important inclusion or exclusion criterion is omitted during the design of a rule).
Build error	The clinical conceptualization and design of the rule were sound, but there was an error building the rule that caused it to malfunction. (eg, if an analyst accidentally enters a wrong value for an age criterion while building a rule).
New code, concept or term introduced, but rule not updated.	A new item was added to one of the data dictionaries used to encode clinical concepts within the EHR's database, but rules that depended on them were not modified to match. This issue manifests in both false negatives (eg, if the pharmacy changes the code for a medication that triggers a drug monitoring alert and the alert stops firing) and false positives (if the laboratory changes the code for the monitoring test, the alert may continue to fire even if the test is done).
Networking	A subtype of "external service issue" added by Thayer et al. (2024). No further description was provided in the source text.
Component failure	A subtype of "external service issue" added by Thayer et al. (2024). No further description was provided in the source text.
Information exchange	A subtype of "external service issue" added by Thayer et al. (2024). No further description was provided in the source text.
Environment migration	Moving the rules from one environment to another (eg, test to production) leads to a malfunction.
New value	The set of allowable values or the meaning of the values of a coded concept is changed.

Organizational policy	A new cause category added by Thayer et al. (2024), not present in the original Wright et al. (2018) taxonomy. Captures malfunctions caused by an organizational or administrative decision rather than a technical defect. (e,g, a CDS malfunction caused by an organizational policy to ignore duplicate alerts within the EHR.)
Inadvertent enabling/disabling	A rule in the production environment is inappropriately enabled or disabled.
Alert text mismatch	The text displayed by the rule and the logic of the rule do not match.
Hardware error	A new cause category added by Thayer et al. (2024), not present in the original Wright et al. (2018) taxonomy. No further description was provided in the source text.
Unaware of component reuse	A “building block” for CDS, such as a criteria specification or a value set, is reused in multiple rules and a knowledge engineer modifies it, intending to change one rule, but inadvertently also causes a change in another rule that uses the same component.
Data source	A new cause category added by Thayer et al. (2024), not present in the original Wright et al. (2018) taxonomy. No further description was provided in the source text.

Table A1.
Full descriptions of malfunction cause categories underlying the classification presented in Table 3, adapted from Wright et al. (2018) and Thayer et al. (2024).