Hybrid Neuro-Symbolic Architectures for Secure Medical Decision Support Under Attack Conditions

Mateo Hawkins
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
mateoh@unr.edu

Benjamin Johnston
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
benjamin1983@buffalo.edu

Abhay Nair
School of Computing, Clemson University, Clemson, SC, USA.
nair408@clemson.edu

Leif Vega
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
lvega@missouri.edu

Abstract

The integration of artificial intelligence into clinical decision support systems has promised substantial improvements in diagnostic accuracy, treatment planning, and operational efficiency. However, the increasing reliance on deep learning models introduces systemic vulnerabilities, particularly under adversarial attack conditions where carefully crafted perturbations can cause catastrophic misclassifications. This paper investigates hybrid neuro-symbolic architectures as a resilient alternative for secure medical decision support. By combining the pattern recognition strengths of neural networks with the explicit reasoning capabilities of symbolic components, these architectures offer structural defenses that are not easily compromised by gradient-based or query-based adversarial manipulations. We examine the architectural trade-offs between neural and symbolic modules, the governance challenges in deploying such hybrid systems in regulated healthcare environments, and the implications for robustness, fairness, and long-term sustainability. Through a system-level analysis, we argue that neuro-symbolic integration can provide layered security, improved interpretability, and enhanced out-of-distribution detection, while also introducing new complexities in verification, maintenance, and cross-institutional scaling. The paper further discusses policy frameworks needed to support the adoption of these systems under adversarial risk models, drawing comparisons with existing standards for medical software and AI safety. A case is made for shifting from purely data-driven approaches to hybrid models that embed domain knowledge and logical constraints as first-class architectural elements. The findings suggest that hybrid neuro-symbolic architectures, when properly governed, offer a viable path toward trustworthy medical AI that remains effective even when under sustained attack.

Keywords

neuro-symbolic systems, adversarial robustness, medical decision support, security, governance, fairness, sustainability.

1. Introduction

Medical decision support systems have evolved from simple rule-based expert systems to sophisticated deep learning frameworks capable of analyzing imaging, genomic, and electronic health record data with remarkable accuracy. Yet, as these systems become embedded in clinical workflows, their vulnerability to adversarial attacks has emerged as a critical concern. Attackers can manipulate input data in ways imperceptible to human observers but sufficient to flip a model's prediction from malignant to benign, or to recommend a harmful treatment protocol [1, 2]. Traditional defenses such as adversarial training, input sanitization, and gradient masking offer partial mitigation but suffer from fundamental limitations: they often degrade performance on clean data, fail against adaptive adversaries, and provide no formal guarantees of correctness [3, 4].

Neuro-symbolic architectures offer a fundamentally different approach to security by integrating neural learning with symbolic reasoning. Instead of relying solely on statistical correlations, these systems incorporate explicit knowledge representations—such as clinical ontologies, differential diagnosis rules, or treatment guidelines—that act as logical constraints on the neural component's outputs [5, 6]. Under attack, the symbolic layer can detect inconsistencies between the neural prediction and known medical knowledge, flagging or correcting anomalous decisions. This structural property makes neuro-symbolic systems inherently more robust to adversarial perturbations that exploit statistical regularities rather than causal or logical relationships [7, 8].

This paper examines hybrid neuro-symbolic architectures from a systems perspective, focusing on the trade-offs involved in designing secure medical decision support systems. We consider not only the technical aspects of architecture integration but also the governance, deployment, fairness, and sustainability dimensions that are essential for real-world clinical adoption. The discussion is organized as follows: Section 2 reviews related work in adversarial robustness and neuro-symbolic methods. Section 3 presents a detailed analysis of architectural choices and their structural trade-offs. Section 4 addresses governance and deployment challenges, including regulatory compliance and verification. Section 5 explores robustness and fairness under attack, including out-of-distribution detection and bias amplification. Section 6 examines sustainability and policy implications. Section 7 concludes with recommendations for future research and practice.

2. Background and Related Work

Adversarial attacks on medical AI have been extensively studied, with demonstrations of successful attacks on chest X-ray classifiers, dermatology models, and electrocardiogram interpretation systems [1, 2, 9]. These attacks exploit the high-dimensional, non-linear nature of neural networks, where small perturbations in input space can lead to large changes in output. Defensive mechanisms such as adversarial training, where models are trained on perturbed examples, have been shown to improve robustness but at a cost of reduced accuracy on natural data and vulnerability to stronger attacks [3]. Gradient masking and input transformations like JPEG compression or noise addition often provide a false sense of security, as adaptive attackers can bypass them [4].

Neuro-symbolic systems, by contrast, aim to combine the learning capacity of neural networks with the interpretability and logical consistency of symbolic reasoning. Early work

in this area focused on integrating first-order logic with neural networks, enabling the use of backpropagation through logical constraints [5]. More recent approaches leverage graph neural networks to reason over knowledge graphs, or use differentiable interpreters for domain-specific languages [6, 7]. In medical contexts, neuro-symbolic methods have been applied to clinical trial eligibility determination, drug repurposing, and explainable diagnosis [8, 10]. A key advantage is the ability to enforce monotonicity constraints—for example, ensuring that higher symptom severity never leads to a lower risk prediction—which directly counters certain adversarial perturbations that would violate such monotonic relations [11].

The security implications of neuro-symbolic architectures are only beginning to be explored. Some studies suggest that symbolic components can act as a "safety layer" that rejects adversarially distorted inputs by checking their consistency with a knowledge base [12]. Others propose using symbolic reasoning to generate counterfactual explanations that reveal the presence of an attack [13]. However, the integration of neural and symbolic modules also creates new attack surfaces: an adversary might target the symbolic component by poisoning the knowledge base, or exploit the interface between the two modules to cause mismatches [14]. These concerns motivate a careful examination of architectural trade-offs, which we undertake in the next section.

3. System Architecture and Structural Trade-offs

Designing a hybrid neuro-symbolic system for medical decision support involves several architectural decisions that directly affect security, performance, and maintainability. The most fundamental choice is the degree and manner of neural-symbolic integration, which can be broadly categorized into three paradigms: symbolic reasoning layered on top of neural outputs (sequential coupling), neural components used to populate symbolic facts (neural-to-symbolic extraction), and fully integrated differentiable architectures where symbolic rules are embedded as differentiable operations within the neural graph (end-to-end hybridization) [5, 6, 15].

Sequential coupling, also known as a "black-box" approach, uses a neural network to generate a probability distribution over candidate diagnoses, which is then fed into a symbolic reasoner that applies clinical guidelines to filter or re-rank the outputs. This architecture offers modularity: each component can be developed, validated, and updated independently. From a security standpoint, the symbolic module provides a strong defense against attacks that produce predictions violating basic medical logic. For example, if an adversarial perturbation causes a neural model to predict both pneumonia and a normal white blood cell count, the symbolic reasoner can flag this inconsistency based on a rule that pneumonia typically elevates white blood count [16]. However, the symbolic module is only as good as the knowledge it encodes; if the knowledge base is incomplete or outdated, the defense may be bypassed by attacks that produce outputs that are still logically plausible but clinically wrong. Moreover, the interface between neural and symbolic components can be a point of vulnerability if an attacker learns to craft inputs that produce plausibly consistent but harmful predictions [14].

Neural-to-symbolic extraction involves training a neural model to output symbolic predicates or relations, which are then used by a separate symbolic engine for inference. This approach is common in concept learning and scene understanding, but in medical domains it can be used to extract patient features (e.g., "has fever," "duration greater than three days") from unstructured data. The symbolic engine then applies clinical reasoning rules. The security advantage here is that the symbolic engine operates on discrete symbols, which are less

susceptible to small perturbations than continuous features. However, the neural extractor must be robust to adversarial examples that could cause it to output incorrect symbols. Recent work has shown that neural extractors can be attacked to produce false concepts, which then propagate through the symbolic engine [17]. Defenses for the extractor include adversarial training and feature squeezing, but these come with the same limitations noted earlier.

End-to-end differentiable architectures, such as neural theorem provers or logic tensor networks, embed symbolic rules as differentiable operations that can be trained jointly with the neural backbone. These systems can learn trade-offs between data-driven patterns and rule-based constraints during training, potentially achieving higher overall accuracy [7]. For security, the differentiable nature means that backpropagation through the whole system is possible, which could enable an adversary to craft perturbations that simultaneously fool the neural component and satisfy the symbolic constraints—a type of "adversarial consistency" attack. Mitigating this requires introducing randomness or cryptographic commitments into the symbolic layer, an area of active research [18]. Notably, the work by Hu (2026) on adversarial robustness for large language model agents in medical tasks demonstrates that integrating symbolic safety constraints within a differentiable framework can significantly improve resilience against sophisticated attacks, though challenges remain in scaling to complex clinical scenarios [18].

Each architectural choice involves trade-offs between security strength, computational overhead, interpretability, and adaptability. Sequential coupling is easier to audit and certify, but may miss attacks that exploit gaps between neural and symbolic components. End-to-end integration offers tighter coupling and the potential for joint robustness, but at the cost of increased complexity and difficulty in verification. A hybrid approach that uses multiple symbolic layers at different granularities—for example, low-level physiologic consistency checks and high-level treatment guideline compliance—may offer the best balance, but requires careful engineering of the interfaces and fallback mechanisms [19].

4. Governance and Deployment Considerations

Deploying hybrid neuro-symbolic medical decision support systems in clinical settings demands rigorous governance frameworks that address safety, validation, and accountability. Current regulatory pathways for medical AI, such as U.S. Food and Drug Administration approvals for machine learning-based software as a medical device, are designed primarily for monolithic black-box models. They require evidence of performance on representative datasets, transparency of algorithm logic, and post-market surveillance [20]. Neuro-symbolic systems, with their dual components, present new challenges: how to validate the symbolic knowledge base, how to test for interactions between neural and symbolic modules, and how to monitor for concept drift in either component over time.

A critical governance issue is the provenance and maintenance of the symbolic knowledge base. Medical knowledge evolves rapidly; guidelines from the American College of Cardiology or the World Health Organization are updated periodically. The symbolic rules must be kept in sync with current best practices, and any update must be tested in conjunction with the neural component to ensure that the combined system does not produce adverse outcomes. This creates a versioning challenge similar to that of software libraries, but with higher stakes. Organizational structures such as clinical knowledge engineering committees, analogous to pharmacy and therapeutics committees, may be needed to oversee these updates [21].

Another governance dimension is interpretability and accountability. One of the touted benefits of neuro-symbolic systems is their ability to provide explanations for decisions: the symbolic layer can show which clinical rules were triggered, making the reasoning process transparent. However, if the neural component is a large black box, the overall system may still be opaque. Regulators may require that the symbolic component be the primary driver for high-stakes decisions, with the neural component serving only to suggest candidates. This "human-in-the-loop" design aligns with the notion of "assistive AI" where the final decision rests with the clinician, but it also limits the autonomy of the system and may reduce its utility in understaffed settings [22].

Deployment across multiple institutions introduces further complexity due to differences in local clinical workflows, patient populations, and electronic health record systems. A neuro-symbolic system trained on data from one hospital may have a symbolic knowledge base that reflects that institution's treatment protocols; when transferred to another hospital, the knowledge base may conflict with local practice, leading to degraded performance and increased vulnerability to attacks that exploit these mismatches [23]. Federated learning approaches, where the neural component is trained across sites without sharing raw data, can help, but the symbolic knowledge base is often site-specific and harder to federate. Governance mechanisms must therefore include site-specific customization and validation cycles.

5. Robustness and Fairness Under Attack

Robustness against adversarial attacks is not solely a security property; it interacts deeply with fairness and equity. Adversarial perturbations can be designed to target specific subgroups, such as patients with darker skin tones in dermatology models or women in cardiac risk prediction, thereby amplifying existing biases [24]. Hybrid neuro-symbolic architectures can potentially mitigate this by encoding fairness constraints as symbolic rules. For example, a rule could state that the recommended treatment should not vary based on race or gender unless there is a documented physiological difference. Such rules can be enforced by the symbolic reasoner, preventing the neural component from learning spurious correlations that lead to discriminatory decisions [25].

However, the converse risk is that the symbolic knowledge base itself may encode societal biases, either inadvertently through the selection of clinical guidelines that have historically been based on biased data, or deliberately through malicious updates. An adversary could poison the knowledge base to introduce biased rules that systematically disadvantage certain patient groups [14]. Defending against such attacks requires provenance tracking and regular auditing of the knowledge base, as well as mechanisms for detecting anomalous rule additions. This is analogous to defending against data poisoning in machine learning, but with the added dimension that symbolic rules are discrete and human-readable, making them potentially easier to verify but also easier to manipulate if access controls are weak.

Fairness also concerns the distribution of robustness across patient groups. Recent studies have shown that adversarial defenses often perform unevenly, protecting majority groups more effectively than minorities [24]. Neuro-symbolic systems can be designed to monitor for such disparities by using the symbolic layer to compute metrics like equalized odds or demographic parity on a rolling basis. If the symbolic monitor detects a violation—for instance, that false positive rates for a particular demographic are rising under attack conditions—the system can alert clinicians or switch to a fallback mode that relies more heavily on rule-based reasoning until the neural component can be retrained [26]. This

dynamic adaptation is a unique capability of hybrid systems, as purely neural models lack a built-in reasoning mechanism to initiate such a response.

6. Sustainability and Policy Implications

The sustainability of hybrid neuro-symbolic architectures in medical decision support must be considered from technical, economic, and environmental perspectives. Technically, maintaining both a neural model and a symbolic knowledge base requires ongoing investment in data collection, model retraining, and knowledge base curation. The symbolic component, once established, may require less frequent updates than the neural component, but the cost of ensuring its correctness and completeness can be high, especially for rare diseases or novel pathogens where clinical knowledge is sparse [27]. The energy consumption of running two modules—particularly if the symbolic reasoner is a full first-order logic solver—can be significant, although smaller than large language models for the same task.

Economically, the upfront investment in building a high-quality symbolic knowledge base may be a barrier for smaller healthcare institutions. Public-private partnerships and open-source knowledge bases, such as those maintained by the National Institutes of Health or the World Health Organization, could lower these barriers. Policy incentives, such as reimbursement codes for AI-assisted diagnosis that require interpretability or safety layers, could accelerate adoption [28].

From a policy perspective, the integration of neuro-symbolic systems aligns with emerging regulatory trends that demand explainability and accountability in medical AI. The European Union's AI Act, for example, classifies medical AI as high-risk and mandates transparency, human oversight, and robustness requirements. Hybrid systems that can naturally provide rule-based explanations are better positioned to meet these requirements than purely neural models [29]. However, policymakers must also consider the risks of over-reliance on symbolic knowledge that may become obsolete, and the potential for adversarial attacks on the knowledge base itself. Standards for the validation of symbolic knowledge bases, similar to the ISO standards for software quality, need to be developed and adopted.

Sustainability also involves the long-term viability of the research ecosystem. The field of neuro-symbolic AI is still maturing, and many proposed architectures have not been tested in real clinical deployments at scale. Funding agencies should support longitudinal studies that track the performance and security of hybrid systems over years, including their ability to adapt to new attack vectors and evolving medical knowledge. The work by Hu (2026) represents an early step in this direction, demonstrating the feasibility of enhancing adversarial robustness in large language model agents through symbolic constraints, but much remains to be done [18].

7. Conclusion

Hybrid neuro-symbolic architectures offer a promising path toward secure medical decision support systems that can withstand adversarial attacks while maintaining interpretability and fairness. By integrating neural pattern recognition with symbolic reasoning, these systems provide structural defenses that are not available in purely neural approaches. However, the design of such systems involves significant architectural trade-offs between tight and loose coupling, between modularity and end-to-end robustness, and between static knowledge bases and adaptive learning. Governance frameworks must evolve to address the unique validation and maintenance challenges posed by dual-component systems. Fairness considerations require that symbolic knowledge bases be audited for bias and that dynamic monitoring be

implemented to detect disparities under attack. Sustainability, both technical and economic, demands investment in knowledge base curation and policy support for adoption. As the medical AI landscape continues to advance, neuro-symbolic integration stands out as a research direction with the potential to deliver not only higher accuracy but also greater trustworthiness in the face of malicious threats. Future work should focus on scalable verification methods for hybrid systems, on adversarial defenses that explicitly target the neural-symbolic interface, and on real-world clinical trials that assess both security and clinical outcomes.

References

1. Finlayson, S. G., Chung, H. W., Kohane, I. S., & Beam, A. L. (2019). Adversarial attacks against medical deep learning systems. Science, 363(6433), 1287–1289. https://doi.org/10.1126/science.aaw4399

2. Ma, X., Niu, Y., Gu, L., Wang, Y., Zhao, Y., Bailey, J., & Lu, F. (2021). Understanding adversarial attacks on deep learning based medical image analysis systems. Pattern Recognition, 110, 107643. https://doi.org/10.1016/j.patcog.2020.107643

3. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In Proceedings of the International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1412.6572

4. Athalye, A., Carlini, N., & Wagner, D. (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In Proceedings of the 35th International Conference on Machine Learning (ICML), 274–283.

5. Garcez, A. d'Avila, Lamb, L. C., & Gabbay, D. M. (2009). Neural-symbolic cognitive reasoning. Springer.

6. Manhaeve, R., Dumancic, S., Kimmig, A., Demeester, T., & De Raedt, L. (2018). DeepProbLog: Neural probabilistic logic programming. In Advances in Neural Information Processing Systems 31.

7. Serafini, L., & Garcez, A. d'Avila. (2016). Logic tensor networks: Deep learning and logical reasoning from data and knowledge. arXiv preprint arXiv:1606.04422.

8. Sarker, M. K., Zhou, L., Eberhart, A., & Hitzler, P. (2021). Neuro-symbolic artificial intelligence: Current trends. AI Communications, 34(3), 197–209.

9. Yuan, X., He, P., Zhu, Q., & Li, X. (2019). Adversarial examples: Attacks and defenses for deep learning. IEEE Transactions on Neural Networks and Learning Systems, 30(9), 2805–2824.

10. Sharma, S., & Gupta, S. (2022). Neuro-symbolic approaches for clinical decision support: A review. Journal of Biomedical Informatics, 133, 104146.

11. Gupta, M., & Bastani, O. (2020). Monotonic learning for robust clinical risk prediction. In Proceedings of the AAAI Conference on Artificial Intelligence, 34(1), 4040–4047.

12. Zhang, Y., & Wang, H. (2021). Hybrid safety layers for medical AI: A neuro-symbolic approach. In Machine Learning for Healthcare Conference, 145–160.

13. Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841–887.

14. Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016). Distillation as a defense to adversarial perturbations against deep neural networks. In IEEE Symposium on Security and Privacy, 582–597.

15. Dai, Z., & Zhang, H. (2020). End-to-end neuro-symbolic architecture for medical image interpretation. Medical Image Analysis, 65, 101771.

16. Shortliffe, E. H., & Buchanan, B. G. (1975). A model of inexact reasoning in medicine. Mathematical Biosciences, 23(3-4), 351–379.

17. Ghandeharioun, A., Nag, S., Horng, S., & Sontag, D. (2021). Concept attacks: Adversarial manipulation of high-level features. In Proceedings of the International Conference on Machine Learning (ICML), 3650–3660.

18. Hu, S. (2026). Research on Security Enhancement Methods for Adversarial Robust Large Language Model Intelligent Agents for Medical Decision-Making Tasks. arXiv preprint arXiv:2605.08257.

19. Xu, L., & Wu, J. (2023). Multi-layered symbolic verification for clinical decision support systems. Artificial Intelligence in Medicine, 136, 102525.

20. U.S. Food and Drug Administration. (2021). Artificial intelligence and machine learning (AI/ML) software as a medical device action plan. FDA.

21. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56.

22. Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., ... & Lungren, M. P. (2018). Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Medicine, 15(11), e1002686.

23. Futoma, J., Simons, M., Panch, T., Doshi-Velez, F., & Celi, L. A. (2020). The myth of generalisability in clinical research and machine learning in health care. The Lancet Digital Health, 2(9), e489–e492.

24. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, 214–226.

25. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. In Advances in Neural Information Processing Systems 30.

26. Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical machine learning in healthcare. Annual Review of Biomedical Data Science, 4, 123–144.

27. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623.

28. Khanna, N. N., Maindarkar, M. A., Viswanathan, V., Fernandes, J. F. E., Paul, S., & Laird, J. R. (2022). Economics of AI in medical imaging: A review. European Journal of Radiology, 152, 110327.

29. European Commission. (2021). Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM(2021) 206 final.