NeuroSymbolic Clinical Decision Agents with Fast Reactive Policies and Slow Diagnostic Reasoning

Lucas Bailey
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
lucasbailey711@colostate.edu

Kiran Mishra
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
mishrakiran@unr.edu

Abstract

The integration of artificial intelligence into clinical decision-making has historically oscillated between purely data-driven deep learning approaches and rule-based symbolic reasoning. Each paradigm offers distinct advantages: neural networks excel at pattern recognition from raw patient data, while symbolic systems provide transparent, logical inference. This paper proposes a hybrid neurosymbolic architecture for clinical decision agents that explicitly separates fast, reactive policies from slow, deliberative diagnostic reasoning, inspired by dual-process cognitive theories. The fast component learns low-latency, high-frequency responses for routine or time-critical tasks such as triage alerts, medication dosing adjustments, and flagging abnormal vital signs. The slow component engages in structured, multi-step diagnostic reasoning using knowledge graphs, patient history, and clinical guidelines to generate explainable differential diagnoses and treatment plans. We examine the structural trade-offs between these two modes, including latency versus accuracy, transparency versus performance, and adaptability versus stability. The paper further explores system-level considerations for deployment in hospital infrastructures, including governance frameworks, data privacy, regulatory compliance, and the need for continuous validation across diverse populations. Challenges of robustness against distributional shifts, fairness across demographic groups, and long-term sustainability of such hybrid systems are analyzed through case illustrations from emergency medicine, chronic disease management, and telemedicine. We conclude with forward-looking perspectives on how neurosymbolic clinical agents can be designed to augment rather than replace human clinicians, ensuring that fast and slow reasoning complement each other within a human-in-the-loop framework. The proposed architecture offers a path toward accountable, adaptable, and trustworthy AI in healthcare.

Keywords

neurosymbolic AI, clinical decision support, dual-process reasoning, reactive policies, diagnostic reasoning, healthcare infrastructure, fairness, robustness.

1. Introduction

Clinical decision-making in modern healthcare is characterized by a tension between the need for rapid responses in acute situations and the demand for thorough, evidence-based diagnostic reasoning. Traditional artificial intelligence systems applied to medicine have largely followed either a neural or a symbolic paradigm, yet neither alone captures the full spectrum of cognitive processes that clinicians exhibit. Neural networks can learn complex

patterns from large datasets but often lack interpretability and robustness to rare events. Symbolic systems, such as rule-based expert systems, offer transparency and logical consistency but struggle to generalize from noisy, high-dimensional data. The concept of dual-process theory, popularized by Kahneman, distinguishes between fast, intuitive thinking and slow, analytical thinking. Translating this cognitive architecture into computational agents has gained traction in robotics and autonomous systems, but its application to clinical decision support remains underexplored. This paper presents a neurosymbolic framework for clinical decision agents that operationalizes fast reactive policies and slow diagnostic reasoning as distinct but interconnected modules. The fast module handles time-critical, low-level tasks that demand immediate responses, while the slow module performs sophisticated, knowledge-guided reasoning to resolve complex cases. We argue that such a bifurcated architecture aligns well with the operational realities of clinical workflows and offers structural advantages in terms of scalability, interpretability, and safety. The discussion centers on system-level trade-offs, deployment challenges, and governance considerations rather than on specific algorithmic implementations.

2. Background and Related Work

The development of clinical decision support systems (CDSS) has evolved over several decades. Early systems such as MYCIN and INTERNIST-I used rule-based reasoning to emulate expert diagnostic processes [1]. These symbolic systems were highly interpretable but brittle when faced with incomplete or noisy data. The advent of machine learning, particularly deep learning, shifted the focus toward data-driven models that could learn directly from electronic health records, medical images, and genomic data [2]. However, these neural models are often criticized as black boxes, lacking the ability to explain their reasoning in clinically meaningful terms. Recent efforts in explainable AI (XAI) attempt to bridge this gap, but post-hoc explanations remain incomplete. Meanwhile, neurosymbolic AI has emerged as a promising hybrid paradigm that combines the pattern recognition capabilities of neural networks with the structured reasoning of symbolic logic [3]. In the context of decision-making, the "fast and slow" framework has been explicitly adopted in reinforcement learning and planning, where a reactive policy is used for immediate actions and a deliberative planner for long-horizon reasoning [4]. The Dsadf framework, for instance, formalizes a dual-process architecture for decision making, integrating fast reactive responses with slow deliberative reasoning in a unified agent [5]. While this work is situated in general AI, its principles are directly applicable to clinical domains. In healthcare, there have been isolated attempts to build hybrid models for tasks such as diagnosis, treatment recommendation, and image analysis, but a cohesive system-level design that separates fast and slow reasoning remains lacking. Moreover, prior work has not sufficiently addressed the infrastructural and governance implications of deploying such agents in real-world clinical settings. This paper aims to fill that gap by proposing a comprehensive architecture and examining its socio-technical ramifications.

3. Architecture of NeuroSymbolic Clinical Decision Agents

The proposed architecture consists of two primary reasoning modules: a fast reactive policy module and a slow diagnostic reasoning module, coordinated by an executive controller. The fast module is designed to operate at sub-second latency, processing streaming data from patient monitors, lab results, and bedside sensors. It uses lightweight neural networks that have been trained on large volumes of time-series data to recognize patterns such as impending sepsis, arrhythmias, or adverse drug reactions. The outputs of the fast module are

immediate action recommendations or alerts, which are executed directly when confidence is high and risk is low. The slow module, in contrast, operates at a longer timescale, often minutes to hours. It relies on a knowledge base comprising medical ontologies, clinical guidelines, drug databases, and patient-specific historical records. The slow module performs multi-step reasoning, such as constructing differential diagnoses by linking symptoms to underlying pathologies, evaluating treatment options against contraindications, and generating comprehensive summaries for clinician review. Both modules share a common patient state representation, but the fast module uses a compressed, low-dimensional representation for speed, while the slow module uses a rich, graph-based representation that supports logical queries. The executive controller arbitrates between the two modules based on factors including urgency, uncertainty, and the novelty of the clinical situation. For routine and stable cases, the fast module may handle most decisions with periodic oversight from the slow module. In ambiguous or high-risk situations, the slow module is triggered to perform deeper analysis, potentially overruling or refining fast recommendations. This architectural separation mimics the cognitive load management observed in experienced clinicians.

4. Fast Reactive Policies in Clinical Settings

Fast reactive policies in clinical decision agents are designed for scenarios where decision latency is paramount. Examples include real-time monitoring of vital signs to detect hemodynamic instability, automated adjustment of insulin infusion rates in intensive care units, and flagging of abnormal laboratory values that require immediate intervention. These policies are typically implemented using deep reinforcement learning or supervised learning on historical episodes, where the action space is constrained to a set of simple, reversible interventions. The training data for such policies must cover a wide range of physiological states and must be carefully curated to avoid bias. One of the key structural trade-offs in fast policies is between sensitivity and specificity. A highly sensitive system may generate excessive false alarms, leading to alert fatigue among clinicians, while a highly specific system may miss critical events. Striking the right balance requires careful calibration using real-world feedback loops. Another trade-off involves the level of autonomy granted to fast policies. In some contexts, such as automated oxygen titration, the policy can execute actions without human approval if the risk is low and the evidence is strong. In other contexts, such as administering a potentially toxic medication, the fast policy should only issue a recommendation pending human verification. The infrastructure for deploying fast policies must handle high-throughput data streams, maintain low latency, and provide robust failover mechanisms if the module fails. Furthermore, the fast module must be continuously validated against distributional shifts, such as changes in patient population, equipment upgrades, or new clinical protocols. The concept of "online learning" within the fast module is controversial due to safety concerns; therefore most implementations rely on periodic retraining with regulatory oversight.

5. Slow Diagnostic Reasoning Mechanisms

The slow diagnostic reasoning module embodies the deliberative, analytical aspect of clinical cognition. It processes a comprehensive set of patient data, including longitudinal history, physical examination findings, imaging reports, and genomic markers, to construct a holistic diagnostic hypothesis. The reasoning engine can be built on top of symbolic knowledge graphs that encode causal relationships between diseases, symptoms, and treatments, often sourced from clinical guidelines and medical literature. The module may utilize probabilistic graphical models or answer set programming to perform abduction, deduction, and causal

inference. One critical advantage of slow reasoning is its ability to generate explanations. For example, the system can trace a chain of reasoning from a presenting symptom to a final diagnosis, citing relevant evidence from the patient record and supporting literature. This explainability is essential for clinician trust and for medico-legal accountability. The slow module also supports "what-if" simulation, allowing clinicians to explore alternative diagnostic paths or treatment scenarios. However, the computational cost of such reasoning is high, and the response time may be incompatible with acute care settings. Therefore, the slow module is typically deployed as an on-demand service, accessed when the fast module flags uncertainty or when a clinician explicitly requests a second opinion. The integration of fast and slow modules presents a coordination challenge: the slow module may need to reinterpret data that the fast module already acted upon, leading to potential inconsistencies. To mitigate this, the executive controller maintains a shared memory of decisions and their contexts, enabling reconciliation. Additionally, the slow module can be used to periodically audit the fast module's performance, detecting drift or systematic errors.

6. Integration and Structural Trade-offs

The dual-module architecture introduces several structural trade-offs that must be carefully managed. The first trade-off is between latency and accuracy. Fast policies sacrifice depth of reasoning to achieve real-time responsiveness, while slow reasoning sacrifices speed for comprehensive analysis. In clinical practice, the appropriate balance depends on the specific use case. For example, in a trauma bay, rapid triage decisions are paramount; in a chronic disease clinic, detailed longitudinal reasoning is more valuable. A second trade-off concerns transparency. Slow reasoning can generate detailed explanations, whereas fast policies, being neural, are inherently opaque. To address this, the fast module can be designed with interpretable features or integrated with post-hoc explanation methods, but these are not perfect. A third trade-off is between adaptability and stability. Fast policies can be updated quickly in response to new data, but frequent updates risk instability and loss of prior knowledge. Slow reasoning, being rule-based, changes slowly as guidelines evolve, providing a stable reference. The challenge is to allow the fast module to adapt to local practice patterns without deviating from evidence-based standards set by the slow module. Governance mechanisms, such as versioned model registries, human-in-the-loop approval workflows, and audit trails, are necessary to manage these trade-offs. Another critical trade-off is computational resource allocation. Running both modules concurrently increases hardware requirements, which may be a limitation in resource-constrained healthcare settings. Cloud-based or edge deployment strategies can mitigate this, but introduce latency and privacy concerns. The architecture thus requires careful engineering of the pipeline, including data preprocessing, model serving, and result fusion.

7. Deployment, Governance, and Infrastructure

Deploying neurosymbolic clinical decision agents in actual hospital environments involves navigating a complex socio-technical landscape. The infrastructure must support continuous data ingestion from heterogeneous sources such as electronic health records (EHRs), medical devices, and laboratory information systems. Interoperability standards like HL7 FHIR are essential for seamless integration, but many institutions still rely on legacy systems. Data privacy regulations, particularly HIPAA in the United States and GDPR in Europe, impose strict constraints on patient data storage, transmission, and processing. Any cloud-based component of the agent must ensure data encryption, access control, and compliance with local laws. Governance frameworks should define clear roles and responsibilities for model

development, validation, deployment, monitoring, and retirement. A clinical safety case must be established, demonstrating that the agent's decisions are at least as accurate and safe as existing practice. Regulatory bodies such as the FDA have begun to issue guidance for AI-based medical devices, including requirements for transparency, continuous learning, and post-market surveillance. The dual-process architecture offers advantages here: the slow reasoning module can serve as a safety monitor for the fast module, flagging potentially harmful recommendations before they are executed. Additionally, the system should incorporate mechanisms for clinician override, logging all interactions for later review. Deployment should be phased, starting with low-risk decision support tasks (e.g., medication reminders) before progressing to higher-risk tasks. The infrastructure must also support periodic recalibration using local data to account for population-specific variations, while guarding against overfitting to small datasets. Sustainability of such systems demands ongoing investment in training data curation, model updates, and IT support, which can be challenging in under-resourced settings.

8. Robustness, Fairness, and Sustainability

Robustness of clinical decision agents is threatened by distributional shifts between training and deployment environments. For instance, a fast policy trained on data from a tertiary care center may perform poorly in a rural clinic with a different patient demographic and different equipment. Adversarial perturbations, even unintentional ones such as sensor drift, can degrade performance. The slow reasoning module, relying on formal knowledge, is less susceptible to such shifts but may fail when its knowledge base is outdated or incomplete. Combining both modules can improve robustness: if the fast module produces an implausible recommendation, the slow module can detect the anomaly and trigger a fallback. Fairness is another critical concern. Data-driven fast policies may encode historical biases related to race, gender, socioeconomic status, or other protected attributes. For example, an algorithm trained on predominantly white populations may underdiagnose conditions in minority groups. The slow module, if its knowledge base is derived from clinical guidelines that are themselves unbiased, could serve as a corrective. However, guidelines can also embed systemic biases. Ongoing monitoring for disparate impact and algorithmic audits using stratified performance metrics are essential. Interventions such as reweighting training data, incorporating fairness constraints, or using counterfactual fairness techniques can be applied to the fast module. Sustainability from an environmental standpoint includes the energy consumption of training and running large neural networks. The fast module, being relatively small, has a lower carbon footprint than the slow module if the slow module relies on large language models or extensive reasoning. Trade-offs between accuracy and energy efficiency must be considered. Social sustainability requires that the system be adoptable by clinicians without causing burnout or deskilling. The fast module should reduce cognitive load, not increase it. The slow module should empower clinicians with insights, not override their judgment. Long-term sustainability also involves economic viability: the cost of developing and maintaining such systems must be justified by improved patient outcomes and reduced healthcare costs.

9. Future Directions

The neurosymbolic clinical decision agent architecture presented here opens several avenues for future research. One promising direction is the development of meta-learned controllers that can dynamically adjust the balance between fast and slow reasoning based on contextual cues, such as patient acuity, workload of the clinical team, and time of day. Another direction is the incorporation of causal inference methods into the slow reasoning module, enabling the

agent to reason about interventions and counterfactuals. This would be particularly valuable for personalized treatment planning. Advances in foundation models and large language models could be integrated into the slow module to provide more natural language understanding and generation, but caution is needed to avoid hallucinations. The fast module could benefit from continual learning techniques that allow safe updates from new data without catastrophic forgetting. From a governance perspective, federated learning could enable multi-institutional training of fast policies while preserving data privacy. The regulatory landscape is evolving; future work should engage with policymakers to create certification pathways for dual-process systems. Human factors research is needed to understand how clinicians interact with agents that sometimes act autonomously and sometimes provide explanations. Finally, cross-domain comparisons with other safety-critical fields such as aviation, autonomous driving, and nuclear power control could yield valuable insights for clinical deployment.

10. Conclusion

This paper has presented a comprehensive architecture for neurosymbolic clinical decision agents that separates fast reactive policies from slow diagnostic reasoning, drawing inspiration from dual-process cognitive theory. The fast module provides real-time, pattern-based responses for routine and urgent tasks, while the slow module offers deliberate, knowledge-driven analysis for complex diagnostic challenges. We have discussed the structural trade-offs between these two modes, including latency, transparency, adaptability, and resource usage. System-level considerations for deployment, governance, robustness, fairness, and sustainability were examined through a socio-technical lens. The proposed framework aims to augment human clinicians rather than replace them, fostering a collaborative human-in-the-loop environment where fast and slow reasoning complement each other. As AI continues to permeate healthcare, architectures that explicitly model the dual nature of clinical cognition will be essential for building trustworthy, accountable, and effective decision support systems. The path forward requires interdisciplinary collaboration among computer scientists, clinicians, ethicists, and policymakers to ensure that these agents serve the best interests of patients and society.

References

1. Shortliffe, E. H., & Buchanan, B. G. (1975). A model of inexact reasoning in medicine. Mathematical Biosciences, 23(3-4), 351–379. https://doi.org/10.1016/0025-5564(75)90047-4

2. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

3. Garcez, A. d., & Lamb, L. C. (2023). Neurosymbolic AI: The 3rd wave. Artificial Intelligence Review, 56(2), 1235–1260. https://doi.org/10.1007/s10462-022-10247-5

4. Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.

5. Dou, Z., Cui, D., Yan, J., Wang, W., Chen, B., Wang, H., ... & Zhang, S. (2025). Dsadf: Thinking fast and slow for decision making. arXiv preprint arXiv:2505.08189.

6. Bates, D. W., & Gawande, A. A. (2003). Improving safety with information technology. New England Journal of Medicine, 348(25), 2526–2534. https://doi.org/10.1056/NEJMsa020847

7. Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28(1), 31–38. https://doi.org/10.1038/s41591-021-01614-0

8. Tjoa, E., & Guan, C. (2020). A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Transactions on Neural Networks and Learning Systems, 32(11), 4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314

9. Marcus, G., & Davis, E. (2019). Rebooting AI: Building artificial intelligence we can trust. Pantheon Books.

10. Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., ... & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484–489. https://doi.org/10.1038/nature16961

11. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9(4), e1312. https://doi.org/10.1002/widm.1312

12. Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—Big data, machine learning, and clinical medicine. New England Journal of Medicine, 375(13), 1216–1219. https://doi.org/10.1056/NEJMp1606181

13. Liu, Y., Chen, P. H. C., Krause, J., & Peng, L. (2019). How to read articles that use machine learning: Users' guides to the medical literature. JAMA, 322(18), 1806–1816. https://doi.org/10.1001/jama.2019.16489

14. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

15. Raji, I. D., & Buolamwini, J. (2019). Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (pp. 429–435). https://doi.org/10.1145/3306618.3314244

16. Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—Addressing ethical challenges. New England Journal of Medicine, 378(11), 981–983. https://doi.org/10.1056/NEJMp1714229

17. Sendak, M. P., Ratliff, W., Sarro, D., Alderton, E., O'Brien, C., O'Connell, J., ... & Fiebig, D. (2020). Real-world integration of a sepsis deep learning technology into routine clinical care: An implementation science approach. JAMA Network Open, 3(4), e203184. https://doi.org/10.1001/jamanetworkopen.2020.3184

18. Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., & Ranganath, R. (2020). Practical guidance on artificial intelligence for health-care data. The Lancet Digital Health, 2(11), e580–e586. https://doi.org/10.1016/S2589-7500(20)30222-9

19. Yu, K. H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. Nature Biomedical Engineering, 2(10), 719–731. https://doi.org/10.1038/s41551-018-0305-z

20. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.