Governing Socio-Technical Risk in an Era of Dual-Use Technology

A Research Paper submitted to the Department of Engineering and Society

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Ethan Rogowsky

Spring 2025

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Advisor

Kent Wayland, Department of Engineering and Society

Introduction: Framing the Sociotechnical Problem

The integration of artificial intelligence (AI) into hospital cybersecurity is reshaping how institutions defend against digital threats. While AI enhances capabilities in threat detection and risk management, it is also being weaponized by attackers to exploit system vulnerabilities. This dual-use nature of AI—serving both institutional defense and adversarial offense—has escalated cybersecurity into a dynamic arms race. Hospitals must now contend with rapidly evolving threats that challenge not just their technical infrastructure but also their organizational readiness, regulatory compliance, and capacity for adaptive governance.

This paper argues that the rise of AI in hospital cybersecurity has catalyzed a shift from static, technical defenses to adaptive sociotechnical governance. In this evolving landscape, cybersecurity is not merely a technical problem but a product of interactions among technologies, institutional actors, and regulatory systems. Understanding how hospitals are responding to AI-driven threats requires analyzing not only the tools they adopt but also the governance frameworks, human networks, and policy pressures that shape their use.

The following sections provide background on AI-enabled hospital security systems, review current literature on AI's dual-use character, and introduce a theoretical framework grounded in Mutual Shaping and Actor-Network Theory. The analysis draws on thematic coding of policy reports and academic sources to explore how hospitals are adapting in response to shifting technological and adversarial conditions.

Literature Review: AI as Defender, Threat, and Governance Catalyst

2.1 AI as a Tool for Cybersecurity Defense

Many studies highlight AI's growing role in enhancing hospital cybersecurity through tools such as Endpoint Detection and Response (EDR) systems, automated anomaly detection, and real-time behavioral monitoring. Hassan et al. (2020) describe how EDR platforms use machine learning to continuously scan for indicators of compromise and automatically respond to threats—reducing dependence on human operators and minimizing the lag between detection and intervention. Similarly, Donepudi (2015) emphasizes AI's ability to learn from evolving threat patterns, identifying intrusions that would escape traditional rule-based systems. These capabilities are particularly valuable in hospital environments, which manage vast and sensitive patient data across complex digital ecosystems.

However, the successful deployment of AI in this domain is not purely technical. Institutions must align AI tools with existing infrastructure, staff expertise, and regulatory requirements—often requiring cross-functional coordination between IT departments, vendors, compliance teams, and clinical administrators. These tensions suggest that even when AI systems are technically available, their effectiveness is shaped by organizational readiness and sociotechnical factors—a point returned to in later theoretical analysis.

2.2 AI as a New Vector of Vulnerability

In parallel, a growing literature documents how attackers increasingly weaponize AI to develop more evasive, automated, and targeted cyberattacks. Raj et al. (2023) and Sharma et al. (2024) identify tactics such as adversarial machine learning, data poisoning, and model inversion, through which attackers manipulate or deceive AI models into misclassifying inputs or suppressing alerts. Poonkuntran (2025) further observes that attackers now test AI systems for behavioral patterns, enabling them to craft tailored inputs that bypass detection algorithms—a practice akin to adversaries "training against" the defenders' models.

These developments were reflected in the 2024 ransomware attack on an Indian healthcare provider, documented in a CyberPeace Foundation (2024) report. In this incident, attackers used machine learning algorithms to scan the hospital's network for vulnerabilities, map administrative controls, and escalate privileges undetected. The attack crippled services and compromised sensitive health data, demonstrating not only AI's effectiveness in adversarial hands but also how rapidly attackers are adapting to the AI-centric security systems being deployed by hospitals.

3.3 Institutional Adaptation and the Sociotechnical Nature of AI Security

The growing sophistication of both AI defenses and AI-enabled attacks has led to a shift in institutional thinking about cybersecurity. Rather than treating AI tools as static solutions, recent literature stresses the need for adaptive, multilayered governance frameworks. Moghadasi (2024), in her doctoral dissertation, proposes a three-layer model—purpose, structure, and function—for assessing institutional readiness to deploy AI securely. This model emphasizes that AI's effectiveness depends on how well it is integrated into an organization's mission, operational routines, and decision-making structures.

Real-world responses to major breaches offer further evidence of this shift. The 2018 SingHealth breach in Singapore, though not caused by AI, catalyzed a national reevaluation of hospital cybersecurity. As documented in the official Committee of Inquiry (2019) report, the attack—enabled by a lack of network segmentation and logging—exposed major institutional vulnerabilities. In its aftermath, Singapore's healthcare system implemented reforms that

included the introduction of AI-based network monitoring tools and more structured oversight frameworks. Here, AI adoption emerged not from technological opportunity alone, but from a process of institutional learning shaped by public accountability, government policy, and perceived risk.

Frameworks such as the NIST AI Risk Management Framework (2023) reinforce this perspective. NIST identifies key principles—transparency, governance, and adaptability—as essential for the responsible deployment of AI in high-risk sectors. Rather than prescribing technical configurations, the framework encourages institutions to view AI security as a continuous process, requiring regular review and context-specific adaptation. This aligns with broader sociotechnical interpretations of cybersecurity as an evolving negotiation between technical capacity, organizational constraint, and external threat.

3.4 Gaps in the Literature

Despite this growing recognition of AI's complexity, relatively few studies bridge empirical case analysis with **sociotechnical theories** of technology and governance. Most works isolate either technical functions or policy recommendations, with limited integration of how these elements co-produce outcomes in real institutional settings. Furthermore, while the SingHealth and Indian ransomware cases are discussed in public and policy literature, they are rarely analyzed in terms of how institutional adaptations reflect broader shifts in governance logic.

This paper contributes to filling that gap by analyzing how hospitals adapt not just to specific threats, but to an AI-driven threat environment that reshapes institutional behavior itself. It applies **Mutual Shaping Theory** and **Actor-Network Theory (ANT)** to interpret these cases not simply as breaches, but as moments of sociotechnical transformation—where technologies, institutions, and adversaries co-evolve.

Theoretical Framework: Mutual Shaping & Actor-Network Theory (ANT)

This analysis is guided by two complementary frameworks from Science and Technology Studies (STS): Mutual Shaping Theory and Actor-Network Theory (ANT). These frameworks support a non-deterministic understanding of how artificial intelligence reshapes hospital cybersecurity—not as a neutral tool acting on a passive environment, but as a technology embedded in and shaped by social, institutional, and material forces.

Mutual Shaping Theory emphasizes the reciprocal relationship between technology and society. Rather than viewing AI as a fixed or self-directing innovation, this approach highlights how hospitals influence—and are influenced by—their implementation of AI. Decisions about when, how, and why to deploy AI systems are shaped by institutional priorities, resource constraints,

regulatory mandates, and evolving threat environments. In turn, these decisions feed back into the organizational structure, altering routines, staffing models, and governance practices.

Actor-Network Theory extends this relational perspective by treating both human and non-human entities—such as AI algorithms, hospital administrators, IT infrastructure, attackers, and policy documents—as actors within a distributed network. ANT draws attention to how security outcomes are negotiated through the interactions and alignments of these heterogeneous actors. Rather than isolating causes in individual agents or technologies, it invites analysis of the complex arrangements that stabilize or destabilize hospital cybersecurity systems.

Together, these frameworks enable a sociotechnical reading of AI-driven hospital security. They guide the analysis by tracing how technologies, threats, and institutional responses co-produce one another across dynamic networks of influence.

Methods

This research is based on a qualitative review of publicly available documents and secondary literature related to the integration of artificial intelligence into hospital cybersecurity. The purpose of the study is not to conduct original empirical analysis, but rather to interpret and synthesize existing materials in order to understand how institutions are responding to the dual-use nature of AI technologies in healthcare security contexts.

The materials examined include peer-reviewed academic articles, government and institutional reports, and media coverage of two widely cited cybersecurity incidents: the 2018 SingHealth data breach in Singapore and the 2024 AI-powered ransomware attack on an Indian healthcare provider. These two cases were selected because they are frequently referenced in scholarly and policy discussions on healthcare cybersecurity and because they offer contrasting examples of institutional responses to digital security threats—one preceding the widespread adoption of AI, and the other involving the direct use of AI in an attack.

Rather than using a formal coding process, the study involved a close reading of selected texts to identify recurring concerns, conceptual framings, and institutional responses to AI-related vulnerabilities. Attention was given to how hospitals and policy bodies characterize the role of AI in both preventing and enabling cyber threats, and how these characterizations reflect broader institutional priorities, constraints, and regulatory pressures. The research is interpretive in nature and draws on Science and Technology Studies (STS) perspectives to consider how sociotechnical systems emerge through the interplay of technologies, organizations, and policy frameworks.

This approach aligns with the theoretical frameworks of Mutual Shaping and Actor-Network Theory, which emphasize the co-production of technological and institutional change. The aim is

not to provide a definitive empirical account, but to situate reported events and documented institutional responses within a broader conceptual understanding of how AI is transforming hospital cybersecurity.

Results and Discussion: Adaptive Governance in a Dual-Use Threat Environment

The synthesis of scholarly literature and institutional reports reveals three interrelated dynamics shaping hospital responses to AI-driven cybersecurity threats: the co-evolution of offensive and defensive AI applications, the emergence of adaptive governance strategies, and the sociotechnical entanglement of institutional actors, technologies, and regulatory environments. These dynamics do not unfold in isolation but are co-produced through interactions between technological innovation, organizational constraint, and adversarial behavior.

6.1 Co-evolution and the AI Security Arms Race

A central pattern in the reviewed materials is the continuous adaptation between attackers and defenders, each responding to the evolving capabilities of the other. Hospitals have increasingly adopted AI tools to enhance detection, automate response, and reduce the burden on human security analysts. However, these same capabilities—such as predictive modeling and anomaly recognition—are being reverse-engineered and exploited by attackers who use AI to map system vulnerabilities, avoid detection, and scale their operations.

This arms race dynamic is especially evident in the 2024 ransomware incident, where adversaries used AI not only to exploit a target system but also to dynamically adjust their attack strategy based on system behavior. Such developments underscore that AI does not provide a stable or enduring advantage. Instead, it accelerates the pace of escalation and forces institutions to think beyond static defenses.

From the perspective of **Mutual Shaping Theory**, this dynamic illustrates how technological innovation in hospital settings is not linear or unilateral. The introduction of AI security tools prompts corresponding changes in attacker behavior, which in turn necessitate further institutional adaptation. Each side's actions reshape the conditions under which the other operates.

6.2 Institutional Adaptation and Governance Flexibility

In response to these shifting threat landscapes, hospitals and governing bodies are adopting more **layered, flexible approaches to cybersecurity governance**. Rather than relying solely on

technical defenses, institutions are integrating AI security tools within broader frameworks that include staff training, vendor oversight, risk audits, and compliance with evolving standards.

The SingHealth breach serves as a turning point in this shift. Although not caused by AI, the attack prompted a regulatory response that emphasized system monitoring, centralized oversight, and the integration of advanced technologies into everyday security routines. The adoption of AI tools in this context did not emerge from technical opportunity alone, but from a broader reevaluation of organizational responsibility and public trust.

This reflects a central insight of **Mutual Shaping**: technologies are embedded in institutional contexts and are shaped by legal, ethical, and operational considerations. AI systems are not simply "plugged in"; their functions are mediated by workforce capacity, budget constraints, and the strategic goals of the organization. Their effectiveness, therefore, depends not only on what the technology can do, but on how institutions are structured to use it.

6.3 Sociotechnical Complexity and Actor-Network Dynamics

Hospital cybersecurity cannot be adequately understood as a conflict between "institutions" and "attackers." Instead, the reviewed materials point to a dense, heterogeneous network of actors—including AI models, human administrators, compliance documents, data protection regulations, and even adversarial algorithms—whose interactions determine security outcomes.

Drawing on **Actor-Network Theory**, this perspective makes visible the complexity of hospital cybersecurity environments. For example, the effectiveness of an AI anomaly detection system depends not only on its code but also on how it is configured by IT staff, monitored by analysts, supported by training programs, and framed within hospital policies. Similarly, adversaries are not just external threats but participants in the network, actively shaping the evolution of defensive systems through their attacks.

Understanding cybersecurity as a sociotechnical system allows for a more nuanced interpretation of institutional behavior. It helps explain why the same AI tool may yield different outcomes in different hospitals, depending on how human and non-human elements are aligned—or misaligned—within a specific network.

6.4 Implications for Adaptive Security Strategy

The reviewed materials suggest that any sustainable cybersecurity strategy must recognize and embrace the **instability and co-evolutionary nature** of AI threats. Attempts to implement fixed solutions or one-size-fits-all technologies are unlikely to succeed in a landscape where attackers learn and adapt as quickly as institutions do.

Instead, hospitals must move toward **adaptive governance models** that integrate technological tools with organizational learning, regulatory responsiveness, and cross-sector collaboration. Frameworks such as the NIST AI RMF emphasize the importance of transparency, adaptability, and participatory oversight—elements that align with the theoretical understanding of security as a sociotechnical process rather than a purely technical challenge.

By reframing cybersecurity as a dynamic network of actors and practices, institutions can begin to anticipate, rather than simply react to, the changing contours of digital threat and defense.

Conclusion

The growing integration of artificial intelligence into hospital cybersecurity systems marks a significant transformation in how digital threats are understood and managed. As this paper has shown, AI's dual role—as both a tool for defense and a vector for attack—has prompted institutions to move beyond static technical solutions and toward more adaptive, sociotechnical forms of governance.

Institutional responses to AI-driven threats are not merely technical upgrades but are shaped by regulatory frameworks, organizational capacity, and interactions with evolving adversarial strategies. Understanding these responses requires viewing hospital cybersecurity as a networked process in which human actors, technologies, and policies are continuously co-producing outcomes.

Future work in this area must continue to explore how healthcare institutions can build resilience not only by adopting advanced technologies, but by cultivating the organizational flexibility and governance structures necessary to navigate an AI-driven security landscape.

References

Committee of Inquiry. (2019). *Public Report of the Committee of Inquiry into the Cyber Attack on Singapore Health Services Private Limited Patient Database*. Ministry of Communications and Information. https://file.go.gov.sg/singhealthcoi.pdf

CyberPeace Foundation. (2024). *AI-powered ransomware attack on a healthcare provider: A research report*. https://www.cyberpeace.org/resources/blogs/research-report-ai-powered-ransomware-attack-on-a-healthcare-provider

Donepudi, P. K. (1970). Crossing Point of Artificial Intelligence in Cybersecurity. Retrieved from https://ideas.repec.org/a/ris/ajotap/0106.html

Hassan, Abeer & Roberts, Lee & Atkins, Jill. (2023). Hassan et al-2020-Business Strategy and the Environment.

Moghadasi, N. (2024). *Enterprise risk management of artificial intelligence in healthcare* (Doctoral dissertation). University of Virginia. https://doi.org/10.18130/63rc-rx48

National Institute of Standards and Technology. (2023). *AI Risk Management Framework (AI RMF 1.0)*. U.S. Department of Commerce. https://www.nist.gov/itl/ai-risk-management-framework

Poonkuntran, S. (2025). *Cybersecurity in healthcare applications*. CRC Press.

Raj, B., Gupta, B. B., Yamaguchi, S., & Gill, S. S. (Eds.). (2023). *AI for big data-based*

engineering applications from security perspectives. CRC Press.

Sharma, N., Srivastava, D., & Sinwar, D. (Eds.). (2024). *Artificial intelligence technology in healthcare: Security and privacy issues*. CRC Press.