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 and Background: Framing the Sociotechnical Problem

The rapid advancement of artificial intelligence (AI) is transforming hospital operations, particularly in the realms of security and patient care. AI technologies are increasingly used to monitor hospital environments, manage sensitive patient data, and support clinical decision-making. As hospitals adopt AI-driven systems for surveillance, diagnostics, and administrative control, their approach to safeguarding both infrastructure and patient trust must evolve accordingly.

Modern hospitals rely on complex digital infrastructures to manage patient data, monitor medical devices, and coordinate care across departments. Protecting these systems involves a combination of technical safeguards—like firewalls, user authentication protocols, and intrusion detection systems—and human processes, such as IT policies, staff training, and compliance auditing. Traditionally, cybersecurity in healthcare has relied on rule-based systems that operate by scanning for known threat signatures, alerting administrators when a suspicious event matches predefined criteria.

However, this rule-based approach has significant limitations. It cannot detect novel or subtle attacks that do not match prior patterns, and it often overwhelms staff with false positives. This is where artificial intelligence enters the picture. But AI does not simply replace human judgment—it redistributes it. These systems must be trained, calibrated, and monitored. Hospitals must decide which data to feed into them, how much authority to grant them, and how to respond when their decisions conflict with human intuition. The effectiveness of AI in cybersecurity, therefore, is not determined by the algorithm alone but by the institutional environment into which it is deployed.

In this way, hospital cybersecurity is best understood as a sociotechnical system—one in which tools, people, workflows, and policies are deeply entangled. A hospital's ability to defend against cyber threats depends as much on its staffing, budgeting, and vendor relationships as it does on the sophistication of its AI. Similarly, regulatory frameworks and compliance obligations shape how and whether these tools are adopted. Understanding AI in this context requires looking not only at what the technology can do, but also at how institutions implement, adapt to, and govern it.

Hospitals are not simply adopting AI—they are struggling to govern it in practice, and that struggle reflects broader tensions between automation, regulation, and trust in the healthcare system. 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.

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

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. This also minimizes the issue of false positives, a case where the system detects an anomaly but it is just normal behavior. 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.

Further, the National Institute of Standards and Technology's AI Risk Management Framework (AI RMF 1.0), released in 2023, provides voluntary guidance for managing AI-related risks. It emphasizes core principles such as transparency, accountability, and ongoing monitoring, encouraging organizations to assess not only technical reliability but also how AI systems align with institutional goals and values (NIST, 2023). This is discussed in further detail in the results, where it is a useful framework for hospitals and medical institutions in an AI landscape.

While the literature reflects a growing recognition of the institutional challenges posed by AI, few studies explore how technologies and governance structures evolve together in practice. This paper uses a framework from Science and Technology Studies to address that gap.

Theoretical Framework: Mutual Shaping and Actor-Network Theory

Mutual Shaping Theory and Actor-Network Theory (ANT) together support a socio-technical view of cybersecurity, emphasizing that outcomes are not determined by technical tools alone, but by how technologies, people, and institutions co-construct systems of action.

Mutual Shaping Theory rejects both technological determinism and purely social explanations, instead proposing that technology and society continuously influence and reshape one another. In the context of hospital cybersecurity, this means that AI tools are not simply "plugged into" existing systems—they are adopted, configured, and governed through complex negotiations involving technical feasibility, regulatory pressure, institutional priorities, and social trust. Likewise, AI itself reshapes hospital structures, prompting shifts in workflow, policy design, and risk management practices.

Actor-Network Theory (ANT) complements this by attending to the heterogeneous actors that form the networks through which security outcomes are produced. ANT treats human and non-human elements—such as AI algorithms, IT teams, attackers, regulatory documents, and risk audit systems—as equally capable of shaping action. From this perspective, hospital cybersecurity emerges from interactions between these diverse entities, whose alignments, conflicts, and translations define the effectiveness of any given security strategy.

These frameworks guide the analysis of the case studies discussed in the methods by making visible the institutional decisions, social forces, and technical infrastructures that co-produce AI adoption and governance. Rather than isolating success or failure in any single tool or actor, this theoretical lens foregrounds the dynamic interplay that shapes hospital cybersecurity in the AI era.

Methods: Interpreting Institutional Response in a Dual-Use AI Landscape

This paper adopts a qualitative, interpretive approach to explore how hospitals and healthcare institutions are responding to the dual-use nature of artificial intelligence in cybersecurity. Rather than conducting original empirical fieldwork, the research draws on existing case documentation, policy reports, and scholarly analyses to understand how institutions frame, implement, and revise AI-based security strategies in the wake of major cyber incidents.

Two case studies serve as focal points for the analysis: the 2018 SingHealth breach in Singapore and the 2024 AI-enabled ransomware attack on a healthcare provider in India. These cases were selected not only because they are widely cited in academic and institutional literature, but also because they illustrate two distinct modes of institutional encounter with AI: one as a post-crisis adoption of AI tools for monitoring and oversight (SingHealth), and the other as a crisis caused by adversarial AI itself (India). Together, they offer a comparative view of how AI enters hospital governance systems—either reactively, through reform, or disruptively, through threat.

The analysis is informed by perspectives from Mutual Shaping Theory and Actor-Network Theory (ANT). These frameworks support a reading of cybersecurity as a sociotechnical process—emerging not only from technical architectures, but from human decisions, institutional logics, and external pressures. Mutual Shaping emphasizes how institutions and technologies co-evolve, while ANT draws attention to how outcomes are produced through networks of heterogeneous actors, including algorithms, staff, protocols, and attackers.

The cases are not analyzed through formal coding but through close reading and comparative interpretation. The aim is to trace patterns in how hospitals make sense of AI's risks and

promises, how they restructure internal practices in response, and what broader socio-technical implications these shifts suggest for healthcare cybersecurity in the AI era.

Case Studies: SingHealth and the India Ransomware Attack

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. Contrary to the above discussion, these tactics from adversarial models can be very hard to detect, and require new technical and social considerations from hospitals. 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.

The 2018 cyberattack on Singapore's SingHealth database represented one of the most serious breaches of patient data in Southeast Asia that used AI, compromising the personal records of 1.5 million individuals, including the Prime Minister. According to the Committee of Inquiry (2019), attackers exploited weak authentication controls and a lack of internal network segmentation, gaining privileged access and maintaining a covert presence for months. Crucially, the breach went undetected not because of a lack of tools, but because existing systems were poorly monitored, logs were not reviewed, and alerts were ignored. The incident exposed deep vulnerabilities not only in technological infrastructure, but in institutional workflows, information governance, and the coordination between IT security teams and hospital leadership.

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.

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. Using their adversarial model, these attackers were able to create an asymmetrical dynamic where the hospital has to constantly be ready to defend against a wide range of unpredictable and evolving threats, while the attackers only need to identify a single overlooked vulnerability or moment of weakness to achieve their objective.

Findings and Discussion

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.

Governance Under Pressure in an AI-Driven Threat Environment

Both case studies illustrate that hospitals tend to adopt or overhaul AI-based cybersecurity tools only in the aftermath of a crisis. In the wake of the SingHealth attack, Singapore's Ministry of Health mandated the development of centralized cybersecurity governance and adopted risk management practices similar to those outlined in the NIST framework. These included structured monitoring protocols, clearer delineation of responsibility across departments, and regular internal audits—each of which shaped how AI tools like anomaly detection software were implemented and evaluated.

Policy frameworks like the NIST AI RMF shape cybersecurity not by prescribing exact technical solutions, but by creating reference points for internal governance. Hospitals draw on these frameworks to justify risk audits, vendor selection criteria, and oversight procedures. This helps

to embed AI systems within broader accountability structures that reflect both technical and social demands.

Despite the presence of frameworks like NIST AI RMF, actual practices of implementation vary widely across institutions. The SingHealth reforms emphasized monitoring and internal accountability, while in the Indian case, there was no evidence of policy-driven resilience at all. This variation reflects the sociotechnical entanglement of AI governance: policies and tools are not universally applied, but shaped by each hospital's internal structure, staffing, political context, and public exposure. These fragmented responses reinforce the idea that AI's integration is shaped as much by institutional capacity and trust as by technical need.

Sociotechnical Complexity and Actor-Network Dynamics

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 Mayo Clinic Platform has acknowledged the utility of the NIST AI RMF in assisting healthcare providers to assess AI trustworthiness, explainability, and bias (Halamka, 2023). This endorsement underscores the framework's relevance and applicability within clinical settings.

For instance, Censinet, a healthcare risk management firm, has introduced an enterprise assessment tool tailored to the NIST AI RMF (Censinet, 2024). This tool enables healthcare organizations to systematically identify, manage, and mitigate AI-related risks, thereby embedding the framework's principles into their operational practices.

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.

Neither success nor failure can be attributed to a single system or decision. The SingHealth breach, for instance, was not caused by the absence of AI, but by failures in configuration, monitoring, and institutional attention. Similarly, in the Indian case, the attackers succeeded not simply because they had advanced tools, but because the hospital lacked coordination between IT, admin, and compliance actors. This reflects Actor-Network Theory's insight: AI tools, IT staff, attackers, regulations, and even audit logs function together in producing (or failing to produce) security. Understanding cybersecurity outcomes requires attention to how these human and non-human actors align—or don't.

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.

Limitations and Future Research

While this study offers insight into how hospitals are navigating the dual-use nature of artificial intelligence in cybersecurity, it is not without limitations. First, the analysis is based on publicly available documents, case reports, and secondary literature rather than original empirical

research. As a result, the findings depend on how these events have been interpreted and framed by institutions and media sources, which may obscure important internal dynamics such as informal decision-making, undocumented policy shifts, or contested interpretations of accountability.

Second, the selection of case studies—SingHealth (2018) and the India ransomware attack (2024)—was useful for illustrative purposes but not exhaustive. While they represent high-profile examples of institutional encounters with AI in cybersecurity, they do not capture the full diversity of healthcare systems, especially in low- and middle-income countries, where both technological capacity and governance infrastructure may differ significantly. Moreover, the specific regulatory environments in Singapore and India may not generalize to contexts such as the United States, the European Union, or smaller health systems with less centralized oversight.

While Mutual Shaping Theory and Actor-Network Theory provide valuable conceptual tools for understanding the sociotechnical nature of cybersecurity governance, they do not offer predictive models or normative guidance. This limits the paper's ability to suggest prescriptive policy interventions, even as it identifies key patterns of institutional behavior and governance strain.

Future research should build on this work by conducting in-depth qualitative fieldwork with hospital cybersecurity teams, administrators, and vendors to better understand how AI-related decisions are made, monitored, and adapted in real time. Comparative studies across national contexts and regulatory regimes would help clarify how different institutional structures shape the uptake and oversight of AI systems. In addition, further integration of ethical, legal, and clinical perspectives would enrich the analysis of how hospitals balance technical innovation with patient trust and public accountability in the context of AI-driven security.

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

Censinet. (2024, August 8). *Censinet delivers enterprise assessment for the NIST Artificial Intelligence Risk Management Framework (AI RMF)*. Censinet. https://www.censinet.com/blog/censinet-delivers-enterprise-assessment-for-the-nist-artificial-intelligence-risk-management-framework-ai-rmf/

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

Halamka, J. (2023, February 14). *NIST provides much-needed AI guardrails*. Mayo Clinic Platform.

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.