Automation Bias and the Future of Clinical Responsibility

Lam, Zhen Yi

Abstract

Automation bias in Clinical Decision Support Systems (CDSS) poses a growing threat to clinical responsibility, especially amid the rise of Artificial Intelligence (AI). While AI can reduce certain errors and improve efficiency, uncritical reliance risks deskilling physicians. Drawing on recent evidence and the Hospital Authority's 2024 report, this essay argues that technological advancements designed to improve efficiency, such as the integration of AI into CDSS, may simultaneously compromise physicians' professional agency through automation bias. This essay proposes mitigation strategies, including technological redesign, algorithmic literacy, and legal reform, to ensure that physicians remain the final moral authority in patient care.

Acknowledgements

The author gratefully acknowledges Professor Joshua Ho for his guidance during the planning phase of this essay.

Introduction

A sixty-year-old male presented with vague abdominal discomfort and slight tachycardia. The physician, exhausted near the end of his shift, entered the data into the hospital's Clinical Decision Support System (CDSS). The algorithm returned a reassuring "Low Risk" score.

This physician hesitated. He noticed a subtle pallor in the patient's face and a slight delay in his responses, signals that sensors had not captured. Yet, he suppressed his intuition, deferring to the machine's statistical aggregate of medical history. Who was he to argue? The patient was discharged with gastritis. Six hours later, the patient returned in septic shock requiring immediate intensive care.

This vignette illustrates a modern medical paradox. Contemporary CDSS have evolved from simple checklists into sophisticated machine learning platforms that act as arbiters of clinical judgement.[1] While reports such as the Hospital Authority's *Safe Care for All* (2024) explicitly recommend improving CDSS to "reduce avoidable adverse events" and "enhance operational efficiency" through developing Artificial Intelligence (AI) capabilities, uncritical implementation risks the atrophy of human expertise.[2] This phenomenon, known as automation bias, represents the human tendency to over-rely on automated cues.[3] This essay explores how automation bias creates a dangerous cycle of deference that amplifies risk, hollows out professional competence, and fractures the bond of trust between patient and physician.

I. The Atrophy of Expertise

Automation bias is rooted in the psychology of the user. It does not stem from malice or laziness; rather, it functions as a coping mechanism that arises when people are confronted with overwhelming amounts of data. For a physician working in an intensive environment, a CDSS providing a diagnosis presents a seductive path of least resistance.[4]

As reliance occurs, the physician's role fundamentally shifts from creating diagnoses to merely approving them. While recent studies confirm that a well-calibrated algorithm catches more than it misses, reducing certain classes of errors,[5] the danger posed by automation bias is qualitative rather than quantitative. By blindly adhering to the generated diagnoses, we dismantle the cognitive infrastructure necessary to catch mistakes, such as the capacity to weigh contradictory evidence.[6] Unlike human errors, which typically follow a logical but flawed trajectory, AI failures are often triggered by statistical correlations invisible to the human eye.[5] When we stop generating our own differential diagnoses, these anomalies become invisible to us.

This reliance triggers "automation-induced complacency."[7] Goddard et al. found that physicians consistently rated abnormal findings as less significant when the CDSS failed to flag them.[7] While the CDSS prevents errors of commission (following wrong advice), it simultaneously induces errors of omission (failing to act when the system does not prompt).[7] The screen becomes the perimeter of the physician's world.

A 2023 study published in *Radiology* vividly quantifies this risk. Researchers examined how radiologists interacted with AI during mammography readings. They found that when AI provided incorrect advice, the diagnostic accuracy of very experienced radiologists plummeted from 82.3% to 45.5% as they deferred to the AI's incorrect judgement.[8] For inexperienced radiologists, the drop was even more severe, falling from 79.7% to a mere 19.8%.[8] This demonstrates how easily over-reliance on automated tools can make physicians abandon their own expertise.

Ultimately, competence demands the constant navigation of uncertainty. It develops by resolving conflicting evidence and synthesizing the narrative of the patient. Perpetual deference to algorithms atrophies this capacity, leading to deskilling.[9] By prioritizing the efficiency of

automated systems, we paradoxically increase the risk of catastrophic failure. As Balytskyi et al. warn, clinical environments are inherently unpredictable; AI systems trained on known pathogens may fail when encountering novel threats, leaving over-reliant physicians unprepared to intervene.[10]

II. The Moral Opacity of Care

Beyond clinical risks, automation bias introduces a crisis of agency. As deep learning neural networks evolve, they become "black boxes." Their logic consists of millions of weighted connections that lead to a diagnosis, a process that is often opaque even to their developers.[11]

Consequently, the concept of informed consent is stripped of its substance. Informed consent relies on the physician's ability to articulate the rationale behind a treatment.[12] If the physician cannot explain why a course of action is recommended, how can a patient provide consent? A patient accepting a treatment solely due to a generated "92% probability of success" is not fully consenting to a medical judgement; rather, it constitutes an act of blind faith. This creates a form of "epistemic injustice" where the patient is reduced to data points rather than treated as a knower of their own experience.[13]

Furthermore, this opacity may also conceal values and assumptions embedded in the algorithm. These systems risk encoding systemic injustice into the architecture of care. Algorithms trained on historical data potentially mirror social biases related to race, gender, and class.[14]

A landmark study by Obermeyer et al. exposed this danger. A widely used healthcare algorithm assigned lower risk scores to Black patients compared to White patients who were equally sick. Since Black patients historically had less access to healthcare (and thus incurred fewer costs), the AI mistakenly concluded they were healthier as the algorithm used healthcare

costs as a proxy for health needs.[15] A physician suffering from automation bias would perpetuate this neglect without realising it.

III. Trust and Identity in Patient Care

The most subtle casualty of automation bias is the relationship between the physician and the patient. Trust in medicine demands more than precision; it demands connection. It is built on the belief that the physician is competent and is acting as a fiduciary in the patient's best interest.[16]

If a patient senses that the physician is merely deferring to a screen, the perception of competence shatters. The patient begins to wonder if the physician is looking at them or the data.[17] When expertise appears to be outsourced, the patient questions the necessity of the human intermediary entirely.

Simultaneously, this dynamic erodes the physician's professional identity. Medicine is a vocation of problem-solving and human connection, yet automation bias reduces physicians to supervisors of algorithms. A study in the *Annals of Internal Medicine* found that physicians spend approximately 16 minutes per patient encounter on Electronic Health Records (EHRs).[18] While this figure predates the integration of advanced CDSS, it signals a trajectory where the patient becomes a peripheral figure in their own consultation.[19]

Ultimately, this creates a form of "moral injury," a concept from military psychology describing the distress of violating one's ethical code.[20] This concept is increasingly relevant in medicine, occurring when physicians feel forced to override their judgement or become bystanders in their own profession.[21] Although the Hospital Authority's 2024 report aims to "reduce the excessive workload of healthcare staff" through integrating AI systems,[2] this strategy risks merely transforming the workload into the high-stakes cognitive vigilance required to

monitor AI.[22] A system that relies on burnt-out, morally injured physicians is a system poised for collapse.

IV. Navigating the Future

We cannot stop the development of automation in medicine; the complexity of modern care requires technology to reduce labour-intensive work and improve operational efficiency.[2] Yet, to prevent the erosion of medical responsibility, we must fundamentally transform physicians' relationship with automation through the aspects of technological design, education, and law.

First, we need a technological redesign that mandates human-centered design principles, introducing "cognitive friction".[23] CDSS should require the physician to input their provisional hypothesis first, then act as a frame of reference to challenge logic rather than replace it.[24, 25] Sutton et al. emphasize that evidence-based mitigation strategies must include systems that force users to actively engage to prevent complacency. Additionally, we must mandate Explainable AI (XAI), algorithms that offer "rationale summaries" that allow the physician to critique the machine's logic.[23, 25, 26]

Simultaneously, we must promote the algorithmic literacy of physicians. Initiatives such as the Artificial Intelligence Literacy course in The University of Hong Kong exemplify this, offering medical students a chance to investigate the mechanisms and development of AI for healthcare. Future physicians must understand how these models work and where they fail, and they must possess the confidence to ignore them.[27, 28] The Hospital Authority's recommendation to "actively [involve] frontline staff in the design and development process" is also a vital step in this direction.[2]

Finally, we must address legal frameworks. Current liability standards incentivise defensive reliance on AI.[29] As long as physicians fear liability for disagreeing with machines, automation bias will persist. Liability standards should adopt a "reasonable reliance" test, where a physician who overrides an AI recommendation should be protected if they can articulate a reasonable clinical rationale.[30] Conversely, blind adherence to an algorithm should not immunise a physician from responsibility for foreseeable harm.

Conclusion

The ultimate challenge is philosophical. If medicine merely processes biological data, we should step aside for the machines. They will eventually be faster and more accurate than we are. However, medicine has never been just a science. It is the art of applying knowledge with empathy to human suffering. By resisting automation bias, we protect the humanity of both the patient and the profession.

While an AI can calculate a survival rate, only a human can understand what it means to survive. To prevent harm and preserve the integrity of the medical profession, we must ensure that AI remains a tool. The physician must remain the final moral authority in care.

References

[1] Berner ES. Clinical Decision Support Systems: Theory and Practice. Cham: Springer International Publishing; 2016.

[2] Hospital Authority. Safe Care for All. 2024.

[3] Van Zyl L. The Unintended Negative Consequences of Artificial Intelligence Use for Psychologists. 2025. https://doi.org/10.31234/osf.io/kzqx3_v1.

[4] Mosier KL, Skitka LJ. Human Decision Makers and Automated Decision Aids: Made for Each Other? ResearchGate 1996;40:201-20.

[5] Saadeh MI, Janhonen J, Beer E, Castelyn C, Hoffman DN. Automation complacency: risks of abdicating medical decision making. AI and Ethics 2025;5:5783-93. https://doi.org/10.1007/s43681-025-00825-2.

[6] Patil SV, Myers CG, Lu-Myers Y. Calibrating AI Reliance-A Physician's Superhuman Dilemma. JAMA Health Forum 2025;6:e250106. https://doi.org/10.1001/jamahealthforum.2025.0106.

[7] Goddard K, Roudsari A, Wyatt JC. Automation bias: a systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association 2012;19:121-7. https://doi.org/10.1136/amiajnl-2011-000089.

[8] Dratsch T, Chen X, Rezazade Mehrizi M, Kloeckner R, Mahringer-Kunz A, Pusken M, et al. Automation Bias in Mammography: The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance. Radiology 2023;307. https://doi.org/10.1148/radiol.222176.

[9] Tytler K. The deskilling dilemma: will clinical AI erode or enhance medical expertise? (UK, 2025) | iatroX Clinical AI Insights. IatroX 2025.

https://www.iatrox.com/blog/clinical-ai-deskilling-evidence-and-strategies-for-uk-doctors-2025 (accessed February 25, 2026).

[10] Balytskyi Y, Kalashnyk N, Hubenko I, Balytska A, McNear K. Enhancing Open-World Bacterial Raman Spectra Identification by Feature Regularization for Improved Resilience against Unknown Classes. arXiv.org 2023. https://arxiv.org/abs/2310.13723 (accessed February 26, 2026).

[11] Sogaard A. On the Opacity of Deep Neural Networks. Canadian Journal of Philosophy 2023;53:224-39. https://doi.org/10.1017/can.2024.1.

[12] Paterick TJ, Carson GV, Allen MC, Paterick TE. Medical Informed Consent: General Considerations for Physicians. Mayo Clinic Proceedings 2008;83:313-9. https://doi.org/10.4065/83.3.313.

[13] Fricker M. Epistemic Injustice: Power and the Ethics of Knowing. 2007. https://doi.org/10.1093/acprof:oso/9780198237907.001.0001.

[14] Franklin G, Stephens R, Piracha M, Tiosano S, Lehouillier F, Koppel R, Elkin PL. The Sociodemographic Biases in Machine Learning Algorithms: A Biomedical Informatics Perspective. Life 2024;14:652. https://www.mdpi.com/2075-1729/14/6/652.

[15] Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. Science 2019;366:447-53. https://doi.org/10.1126/science.aax2342.

[16] Ludewigs S, Narchi J, Kiefer L, Winkler EC. Ethics of the fiduciary relationship between patient and physician: the case of informed consent. Journal of Medical Ethics 2022;51:59-66. https://doi.org/10.1136/jme-2022-108539.

[17] Shan W. Digital health's impact on the patient-doctor relationship in a primary healthcare setting: A qualitative study. Annals Singapore 2025. https://annals.edu.sg/digital-healths-impact-on-the-patient-doctor-relationship-in-a-primary-healthcare-setting-a-qualitative-study/ (accessed February 25, 2026).

[18] Overhage JM, McCallie D. Physician Time Spent Using the Electronic Health Record During Outpatient Encounters. Annals of Internal Medicine 2020;172:169-74. https://doi.org/10.7326/m18-3684.

[19] Sharko M, Cole CL. Integrating Artificial Intelligence Support in Patient Care While Respecting Ethical Principles. JAMA Network Open 2025;8:e250462. https://doi.org/10.1001/jamanetworkopen.2025.0462.

[20] Litz BT, Stein N, Delaney E, Lebowitz L, Nash WP, Silva C, et al. Moral injury and moral repair in war veterans: A preliminary model and intervention strategy. Clinical Psychology Review 2009;29:695-706. https://doi.org/10.1016/j.cpr.2009.07.003.

[21] McGoldrick KE. If I Betray These Words: Moral Injury in Medicine and Why It's So Hard for Clinicians to Put Patients First. Anesthesiology 2024;141:1219-21. https://doi.org/10.1097/aln.0000000000005166.

[22] Mewborn EK, Fingerhood ML, Johanson L, Hughes V. Examining moral injury in clinical practice: A narrative literature review. Nursing Ethics 2023;30:960-74. https://doi.org/10.1177/09697330231164762.

[23] Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. npj Digital Medicine 2020;3. https://doi.org/10.1038/s41746-020-0221-y.

[24] Eapen B. Towards a theory of adoption and design for clinical decision support systems. McMaster University 2021. https://macsphere.mcmaster.ca/items/ba07ef2e-e26f-47d8-92cc-3df20d3906fa (accessed February 26, 2026).

[25] Chang T-M, Kao H-Y, Wu J-H, Su Y-F. Improving physicians' performance with a stroke CDSS: A cognitive fit design approach. Computers in Human Behavior 2016;54:577-86. https://doi.org/10.1016/j.chb.2015.07.054.

[26] Roychowdhury S, Lanfranchi V, Mazumdar S. Evaluating explanation performance for clinical decision support systems for non-imaging data: A systematic literature review. Computers in Biology and Medicine 2025;197:110944. https://doi.org/10.1016/j.compbiomed.2025.110944.

[27] Schwartzstein RM. Clinical Reasoning and Artificial Intelligence: Can AI really think? Transactions of the American Clinical and Climatological Association 2024;134:133.

[28] Yoganathan V. AI threat of "de-skilling" for medical trainees. Juta MedicalBrief 2025. https://www.medicalbrief.co.za/ai-threat-of-de-skilling-for-medical-trainees/ (accessed February 25, 2026).

[29] Kellar R. AI in Healthcare: Redefining Liability for Doctors and Hospitals. British Journal of Hospital Medicine 2025:1-6. https://doi.org/10.12968/hmed.2025.0212.

[30] Bartlett B. Taking Responsibility for AI Systems in Healthcare. University of Manchester 2025. https://research.manchester.ac.uk/en/studentTheses/taking-responsibility-for-ai-systems-in-healthcare/ (accessed February 25, 2026).