Navigating the Risks of Artificial Intelligence Foundation Models in Healthcare: How Health Systems Can Respond

Warren Poquiz [1]

Abstract: Foundation Models (FMs) have unveiled a new phase in the Artificial Intelligence (AI) era, characterized by significantly larger datasets and massive computational power. This analysis examines the applicability of FMs in the healthcare sector and how their advanced functionalities, such as in-context learning, can enhance overall organizational performance by increasing efficiency, accuracy, and predictability. However, scholarly works over the past decade have primarily focused on the implications of AI's pervasive application in society, and there remains a critical need to deepen the discussion on AI governance, particularly in the healthcare domain. The rapid advancement of AI models, combined with insufficient regulatory oversight, poses significant risks to patients and Healthcare Organizations (HCOs), including privacy breaches, adversarial attacks, model opacity, and algorithmic biases. To address these risks, this paper calls for the promotion of a three-layer governance structure for HCOs based on the hourglass model for AI governance by Mäntymäki et al. (2022).

Introduction

The rise of Generative Artificial Intelligence (AI) is quickly shaping the Fourth Industrial Revolution. AI's computational prowess and advanced algorithmic capabilities have ushered in an era never seen in history while presenting new challenges in navigating the intricate play between machine intelligence and human existence. Recently, there has been a surge in popularity with the use of Large Language Models (LLMs) such as the Generative Pre-Trained Transformer (GPT). LLM is a significant advancement in Natural Language Processing (NLP), a subset of AI explicitly focusing on a computer's ability to comprehend text and spoken words like humans. NLP has revolutionized digital technology through its contributions, such as chatbots, virtual assistants, and language translation. However, the emergence of LLMs and their ability to train on large amounts of data significantly enhances current NLP features by providing contextually relevant texts based on memory. In healthcare, the profound benefits of these models offer an innovative solution to longstanding problems in care delivery.

Notably, the rapid rate at which technological innovations have permeated society characterizes the reduction in the lag of momentous technological advancements in the

[1] Corresponding author. Health Care Administration Program, Texas Woman's University. wpoquiz@twu.edu

twentieth century. For instance, it took more than 200 years from when the steam engine was developed to when Henry Ford built the first car, while it only took less than 50 years from the first call on a wireless handheld device to the development of smartphones with embedded AI technology (Makridakis, 2017). The same pattern can be observed in ChatGPT's latest LLM GPT-4 release, less than a year after its previous groundbreaking iteration (GPT-3) went public in 2022. Subsequently, many renowned names who hold pragmatic views of the technology, including Tesla CEO Elon Musk, warn about AI's "profound risk to society and humanity" and call for a halt on AI training for at least six months (Future of Life Institute, 2023). Geoffrey Hinton, widely known for his works in deep learning and neural networks, also warns about the dangers of AI and calls for urgent investment in AI safety and control (Kleinman & Vallance, 2023). This significant challenge in technological shifts mirrors the inability of governance initiatives to keep up with rapid innovative advancements. For this reason, the World Economic Forum published a white paper articulating that reliance on government legislation regarding rapidly advancing technology is ill-advised as it is likely to be outdated before implementation (2016).

Over the past decade, numerous studies have predicted and outlined the effects of widespread AI use in society; however, the specific focus on AI governance needs to be expanded in the literature, especially in healthcare. This essay will answer two fundamental research questions: What are the inherent risks of an AI-driven healthcare organization (HCO), and how can HCOs appropriately respond to these risks? The paper will identify four potential implementation risks associated with Foundation Models (FMs) in the healthcare landscape and call to promote a three-layer governance structure guided by the principles of ethical AI and applicable regulations.

Foundation Models: The Key To AI-Driven HCOs

"Foundation Models" or FMs is a term coined in 2021 by the Stanford Institute for Human-Centered Artificial Intelligence (HAI) (Bommasani et al.,2021). Bommasani et al. define the term as "any model that can be trained on broad data," adapted, or fine-tuned to a wide range of downstream tasks. The Center for Research on Foundation Models (CRFM) simplifies the definition: train a single model on a vast dataset and customize it for various applications (n.d.). Notable examples of FMs in deployment include LLMs like GPT-3. While LLMs are tasked explicitly with generating and interpreting human-like texts, FMs generally have a broader application by integrating multiple modalities (text, images, videos, etc.) across different models with specific tasks or purposes. This training process is illustrated below:

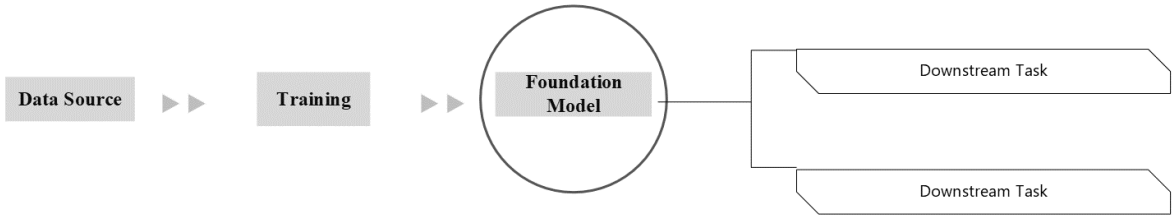

Fig.1 Foundation Model Framework

FMs are rooted in the principles of Artificial Neural Networks (ANNs) and Self-Supervised Learning (SSL), both concepts that have existed for decades (Bommasani et al., 2021). ANNs are systems consisting of artificial neurons, organized in layers, that mirror the behaviors and functions of the human brain. On the other hand, SSL is a type of learning wherein the data "supervises itself for training the model" and instructs its network on what is right or wrong (Rani et al., 2023, p. 2761). The principle behind SSL was derived from how infants learn through observation with little interaction with their surroundings (Rani et al., 2023). Further, since SSL works on unlabeled data, it virtually eliminates the time-consuming and often costly manual annotation of data. However, what makes FMs so fundamentally powerful compared to other AI models is their ability to simultaneously apply the principles of ANNs and SSL at a much larger scale, often measured in parameters. A model's parameter correlates with its ability to discern complex patterns from the data; thus, the greater the parameters, the more it yields superior outputs. For example, GPT-3 has a scale of 175 billion parameters or nearly 45 terabytes of text data (Broadhead, 2023). While training data for GPT-3 is currently not publicized, it is estimated that the model was trained on 500 billion words from the internet (The Alan Turing Institute, 2023). GPT-3's previous iteration (GPT-2), released in 2019, only had 1.5 billion parameters.

The swift progress in scale can be linked to the exponential growth of computational power, also known as "compute," accessible for training datasets. The compute consumption in LLMs like GPT-3 is measured in petaFLOPS-days—the number of computations performed in one day by a computer calculating a thousand trillion computations per second (Power, 2022). GPT-3 required 3,640 petaFLOPS-days to train. A standard laptop would take several thousand years to reach the same number of computations used in training GPT-3. A 2012 paper highlighting an image classification architecture popularly known as "AlexNet" demonstrated how increased computing power can lead to superior results (Krizhevsky et al.,2012). The model in the study outperformed human-level accuracy in image recognition by simply increasing computing power in training a convolutional neural network. These findings led researchers to believe that increasing compute in training top models would lead to better performances, subsequently resulting in a significant rise in computing demands (Power, 2022). From 1959 to 2012, computing power generally doubled every two years; however, since the 2012 study, computing power has doubled every three and a half months (OpenAI, 2018).

In healthcare, applications of AI models have historically been isolated to high-level predictive capabilities of Deep Learning (DL) algorithms for single-purpose tasks such as enhancing image analysis to recognize potentially cancerous lesions in radiology (Fakoor et al., 2013) or risk scoring models such as predicting congestive heart failure (CHF) and chronic obstructive pulmonary disease (COPD) based on clinical data (Cheng et al., 2016). With the advent of FMs, the applications of AI in healthcare now also include advanced functionalities such as in-context learning—the ability to learn from a few examples in the context through analogy (Dong et al., 2022). Fig. 2 visualizes the application of an FM in a healthcare organization. The data is extracted from multidisciplinary sources in care delivery that include both clinical and non-clinical stakeholders. The data gathered will generate multimodal inputs such as clinical notes, diagnostic history, or key performance indicators (KPIs), including financial and operational margins. The foundation model will be

trained on this data to be applied to several downstream tasks in the health system, such as personalized medicine and context-based chatbots for the patient, efficient assistive tools in diagnosis and treatment for the providers, and analytics dashboards that will aid administrators in making informed decisions based on accurate real-time data across various organizational functions. Previous research has also proposed a comprehensive application of DL techniques in healthcare organizations similar to the functionalities of an FM (Miotto et al., 2018). However, the study suggested models that must be constantly updated to follow the changes in patient populations, which can be labor-intensive and expensive. FMs do not focus on specific tasks as they capture a wide range of knowledge from broad organizational data, thereby eliminating the need to train other models in the system from scratch.

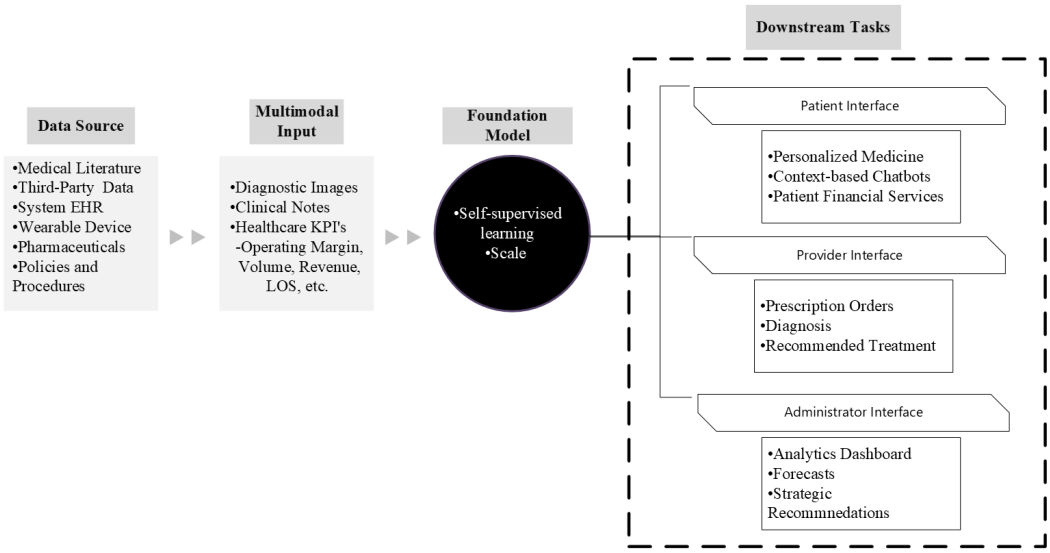

Fig.2 Foundation Model Application in Healthcare (Adapted from Bommasani et al.,2021)

Risks

As FM discussions continue to integrate into healthcare, it becomes imperative to understand the inherent risks posed by implementing them in the healthcare workflow, including privacy, security, explainability, and fairness.

The HIPAA Privacy Rule in the Age of AI

Given the magnitude of the datasets required to train AI systems, it is no surprise that the safeguarding and privacy of data have constituted focal points in most AI legal challenges. AI's hunger for massive amounts of information and healthcare's highly regulated landscape will make it challenging to coordinate the exchange of health information between HCOs and AI developers. In a 2019 class action lawsuit, a patient sued Google and the University of Chicago Medical Center for alleged disclosure of medical information of nearly every patient from the hospital system without removing detailed time stamps and clinical notes. Google assured that data were de-identified, which the plaintiff claimed to be

highly misleading, citing Google's tremendous data mining capabilities make them "uniquely able to determine the identity of almost every medical record the University released" (Dinnerstein v. Google, 2019). *Dinnerstein v. Google* raised questions about whether complete anonymization of data can be actually achieved, especially with the cross-linking capabilities of modern technology. Previous research has demonstrated compromised anonymity in genomic studies where anonymous participants can be identified by analyzing Y-chromosome sequences from public genealogy websites containing their distant relatives' surnames (Gymrek et al., 2013). Another study evaluated an algorithm's ability to re-identify thousands of physical activity data in wearable devices that have de-identified health information and found that the algorithm successfully re-identified more than 80% of the demographic (Na et al., 2018). The *Dinnerstein* case suggests that current anonymization practices do not prevent large digital companies from cross-linking geographical coordinates of Google users and their exact dates and times of arrival and departure from specific locations to timestamps in the health record, identifying anonymous patients by name, physical and email addresses, duration of encounter, etc. (Dinnerstein v. Google, 2019).

The Health Insurance Portability and Accountability Act (HIPAA) authorizes the disclosure of de-identified medical records by third parties as long as there is a low risk that information could be used "by an anticipated recipient to identify an individual who is a subject of the information" (Standards for Privacy of Individually Identifiable Health Information, 2000). However, technological progress at the time of the rule's passing significantly pales in comparison to the vast scale of technology adoption we are witnessing today. Cohen and Mello discussed the implications of the outdated privacy law and its ineffectiveness in addressing data challenges, calling for a reassessment of data-sharing governance in the 21st century (2019). Data experts have also proposed techniques to virtually eliminate privacy risks, such as using synthetic data with simulated datasets (Gaffney, 2023), while others have taken a much broader approach, shifting the discussion toward data ownership by analyzing patient health information within the intellectual property framework (Liddell et al., 2021).

Security Risks: Adversarial Reprogramming, Overlearning, and Centralization

The vulnerabilities associated with FMs extend far beyond data-related concerns. Security threats can emerge from adversarial access to the model itself. As advanced technology progresses, it also brings about a continued evolution of cybersecurity attacks, frequently targeting high-value subjects such as the healthcare industry through ransomware (Kiser & Maniam, 2021). However, the broad spectrum of AI capabilities introduces new pathways for cyber threats to infiltrate systems that can directly affect the clinical workflow. Thus, the deployment of FMs in HCOs and the healthcare industry must be thoroughly assessed by administrators and regulatory leaders, with a specific emphasis on the unique clinical harm they pose to patients. Due to its infancy, the limited literature on FMs has yet to uncover its full potential, rendering any current deployment more akin to prototypes rather than fully-developed implementations.

One common security flaw in AI models is adversarial reprogramming, where a model is repurposed to perform a new task chosen by an attacker, even if the model was not trained for the task (Elsayed et al., 2018). These attacks are incredibly parasitic in nature as they influence a model's functionality rather than its hardcoded output. For instance, Chu et al. outlined the potential dangers of external adversarial networks that can artificially modify imaging results (output) in radiology (2020). In adversarial reprogramming, which is naturally internal, an attacker would not have to modify an output since the model itself has been repurposed to produce flawed imaging results (such as modifying the lesion size, location, etc.) without the knowledge of its developers or users. This has tremendous implications for clinical decision-making as attacks could potentially result in misdiagnoses of abnormalities and life-threatening conditions.

Another security threat to FMs is overlearning. Song and Shmatikov define the term as a phenomenon where "representations learned by deep models when training for seemingly simple objectives reveal privacy- and bias-sensitive attributes that are not part of the specified objective" (2020). In a healthcare implementation, overlearning specifically concerns the amount of sensitive information that can be accessed or disclosed by covered entities under HIPAA. While the Privacy Rule allows the disclosure and use of health information, it also effectively excludes records that are subject to the Family Educational Rights and Privacy Act (FERPA), including "employment records that a covered entity maintains in its capacity as an employer and [an educational institution]" (Standards for Privacy of Individually Identifiable Health Information, 2000). The overlearning tendency of FMs can potentially de-censor these excluded records by enabling the recognition of sensitive information even if it is not present in the training data. Song and Shmatikov highlighted the inadequacy of privacy protection technologies and the regulations that govern them since there are currently no known techniques to censor these "overlearned attributes" (2020).

This analysis has previously discussed the ability of FMs to homogenize the methodologies adapted to downstream applications. Consequently, this inherent centralization can also represent a single point of failure for all downstream tasks (Bommasani et al., 2021). In essence, previously discussed privacy and security risks where adversaries influence either the model or the data can impact not only one single-purpose task but all downstream tasks in the model. Carlini et al. found that LLMs that have been trained on private datasets can be infiltrated by adversaries to extract private information (2021). This means that FMs that are trained on organizational data run the risk of exposing their private data on all downstream applications for adversarial attacks, including model stealing. A more recent and prominent example of such an incident is the model leak of Facebook's "LLaMa" (Large Language Model Meta AI) in early 2023 (Cox).

Interpretable AI and Clinician Trust

The intricate internal workings of AI models have frequently led them to be widely considered as black box models. A black box model can be either a function that is too complex for human intelligence to comprehend or a function that is proprietary (Rudin, 2019). The ability of FMs to train on a vast amount of complex data enables them to

potentially "do unforeseen tasks and do these tasks in unforeseen ways" (Bommasani et al., 2021, p. 123), making them extremely opaque. Further, the predominant focus of interpretability methodologies and initiatives for AI on single-purpose models presents a notable challenge in achieving explainability on FMs because FMs are models influencing an array of other downstream models. In healthcare, the ability to explain and interpret FMs will be critical for user acceptance, trust, and practice of evidence-based medicine. For instance, one study found that the ability to explain and interpret the decision-making process of AI-driven models significantly impacts a physician's behavior towards AI, particularly their trust in the model and intent to use the technology (Liu et al., 2022). Another study highlights the impact of unexplainable models on patient-centeredness, implicating that opaque algorithms can effectively demote patients to "passive spectators in the medical decision-making process" (Amann et al., 2020, p. 8).

While no significant laws currently govern AI in the United States, the General Data Protection Regulation (GDPR) passed by the European Union (EU) in 2018 includes a right-to-explanation provision making it obligatory to explain an algorithm's decision-making process (European Union, 2016). Most major US-based tech corporations must comply with this law as long as they have EU-based consumers, hence the recent emergence of cookie pop-ups on websites asking for consent to collect information. Similarly, the White House Office of Science and Technology Policy (OSTP) released an "AI Bill Of Rights" in 2022 outlining five principles associated with the proper deployment of AI, including explaining an AI system's functionalities in plain language. In 2023, the most comprehensive AI law was effectively passed in the EU—the AI Act. The law aims to address the risks associated with AI without constraining technological development.

One provision of the AI Act allows developers access to "high-quality datasets within their respective fields" (European Union, 2022, p. 29). Enacting a similar law in the U.S. would pose difficulties due to existing privacy regulations within HIPAA. Consequently, Bak et al. predict the possibility of an overall AI ban in healthcare if developers cannot access private health information to test and explain models (2022). These significant regulatory movements indicate that the interpretability and explainability of AI systems will be an integral part of the ongoing discussion toward a comprehensive AI governance structure west of the Atlantic.

AI and Equitable Care

Fairness and bias in algorithms are central concerns in the development and implementation of AI models. Since most models are trained on real-world data, they often reflect inherent societal inequities. Numerous research studies have identified widespread biases in many algorithms deployed in different sectors and functions, including the criminal justice system (Van Dijck, 2022), child protective services (Keddell, 2019), and human resources (Tuffaha, 2023). In healthcare, a 2019 study focusing on racial bias in an algorithm found that black patients were identified to be at a much lower risk than white patients despite being in the same sickness level (Obermeyer et al., 2019). The study also found that the algorithm had assigned risk scores based on health expenditures accrued, which can be misleading if one group has substantially lower access and, thus, lower

utilization and spending. Further, the study found that risk scores for black patients would more than double if biases were removed.

In an AI-driven HCO, the provision of equitable healthcare may very well depend upon the leaders' and developers' understanding of systemic disparities in diverse patient populations. A *Futurescan* survey of healthcare executives found that only 12% of health systems fully understand the profiles of their patient populations (2023). Understanding the patient population's economic, social, racial, and cultural backgrounds will be crucial in identifying algorithmic biases in future healthcare FMs.

Developing a Robust AI Governance Program

The *2023 Futurescan* survey results on healthcare trends indicate that 28% of health systems anticipate being prepared to adopt systemwide AI models by 2028 to manage care delivery. However, while the regulatory landscape of AI remains unclear, the responsibility falls on HCOs to establish a robust organizational governance structure to oversee the development and implementation of the technology. This essay is a call to promote the use of the governance framework illustrated in Fig. 3 in HCOs, based on the Hourglass Model of AI Governance by Mäntymäki et al. (2022). The model depicted has been slightly adjusted to accommodate the distinct characteristics of a health system. The model consists of three fundamental layers: environmental, organizational, and operational/AI system layer.

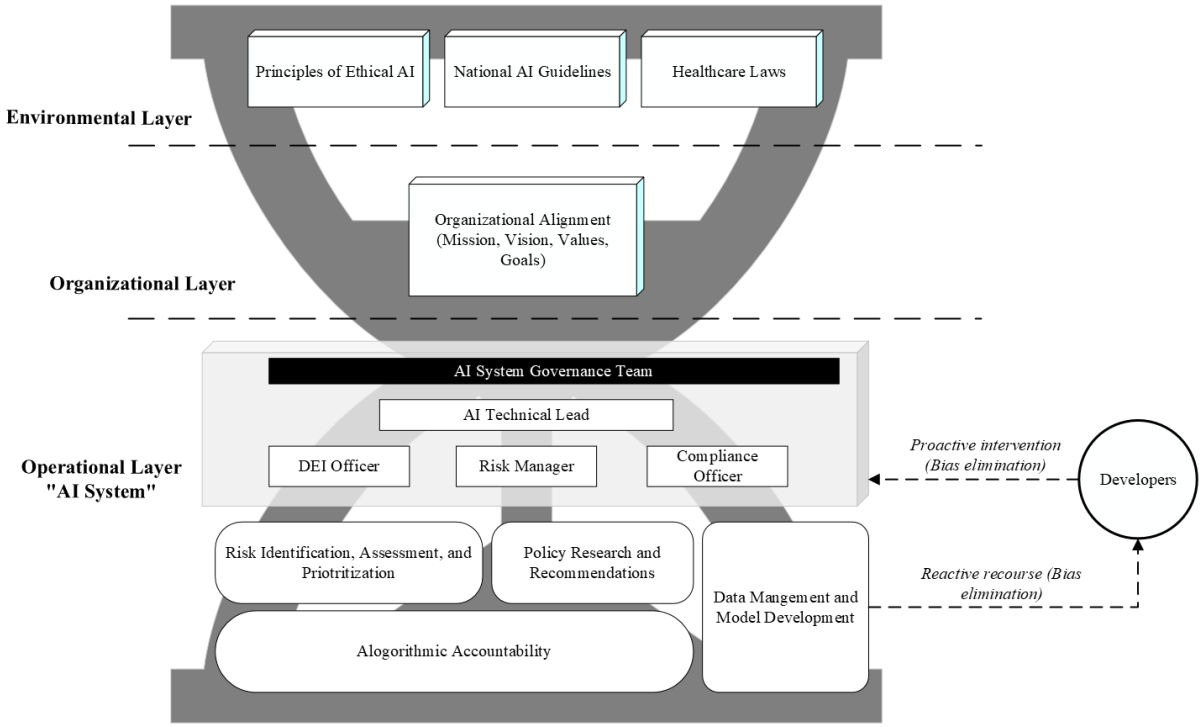

Fig. 3 FM Governance Structure

Environmental Layer

Mäntymäki et al. define this layer as an organization's 'contextual environment' (2022). Since there is no comprehensive legal framework for AI in the U.S., healthcare laws such as HIPAA constitute the most binding regulation within the environmental layer of an AI-driven HCO. This layer also encompasses the ethical principles of AI that will guide the organization. Without hard AI laws, having an ethical framework that clearly defines the appropriate and inappropriate use of the technology is highly crucial. Floridi and Cowls identified an overarching framework for ethical AI, incorporating the four traditional principles of bioethics (beneficence, non-maleficence, autonomy, and justice), along with the addition of a fifth principle: explicability. Explicability aims to comprehend and hold accountable the decision-making processes of an AI model (Floridi & Cowls, 2021).

Organizational Layer: Strategic Alignment

The organizational layer details the HCO's strategic AI initiative with a specific focus on the problem or opportunity that the technology is supposed to address. The strategic AI initiative must also align with the organization's mission, vision, values, and goals and include a specific plan with detailed timelines and meaningful success measures. This strategic alignment ensures that the AI system will perform according to its intended purpose.

Comprehensive Strategic Planning

This technological venture involves defining clear, actionable objectives accompanied by specific, measurable outcomes. Through a comprehensive needs assessment, the strategic team must construct a roadmap that is realistic and attainable, clearly identifying not only the "what" and "why" but also the "how" and "when" of AI deployment. The organizational layer must also foster cross-departmental collaboration to ensure AI initiatives are well integrated across all facets of the HCO, from clinical care to administrative functions, ensuring that AI tools are developed and implemented with a holistic view of the organization's needs, promoting synergies between departments and avoiding siloed efforts.

Operational Layer: The AI System

The operational layer or AI system is the bottom layer in the governance framework, which includes the core AI governance team. The governance team will be led by an AI executive, a leader who possesses specialized knowledge and expertise on the foundation model. In addition, officers or representatives from Risk Management, Compliance and Accreditation, and Diversity, Equity, and Inclusion (DEI) must comprise the rest of the core team. The core

governance team will play a pivotal role in managing and monitoring the AI system throughout its lifecycle.

Core Operational Functions

The core governance team will oversee several key functions essential in the successful deployment and management of AI systems.

> **Risk Identification, Assessment, and Prioritization.** This involves continuous monitoring of potential risks that AI systems may pose in both clinical and non-clinical functions, from patient care to privacy and ethical concerns, and prioritizing them based on severity and likelihood.
>
> **Policy Research and Recommendations.** The core team will diligently monitor the changing landscape of AI guidelines, regulations, and best practices. The team will also be tasked with formulating policy recommendations that align with national standards and organizational objectives.
>
> **Algorithmic Accountability.** The team will ensure that the AI system operates transparently and accountably, with a mechanism in place to review and audit AI-driven decisions.

Developer Engagement

A critical aspect of the operational layer's functions is its interaction with AI developers. This two-way interaction involves working with developers to proactively identify and eliminate biases within the system before they impact patient care and operations. Should biases be detected post-implementation, the operational layer coordinates with developers to address and resolve these issues swiftly.

Conclusion

Foundation Models offer innovative solutions to longstanding healthcare problems in clinical and nonclinical functions, potentially optimizing an HCO's overall organizational performance. However, the understanding of this technology's potential impact on healthcare operations and patient care remains limited. Due to the lack of comprehensive regulatory oversight, HCOs must meticulously approach the adoption of FMs, which must be paired with a robust organizational governance structure and a core governance team to ensure trustworthy, ethical, and patient-centered AI use.

The risks addressed in the integration of FMs in healthcare primarily include privacy breaches, security vulnerabilities, model opacity, and algorithmic biases. These risks encompass the potential for unauthorized access to sensitive data, manipulation of AI systems by malicious actors, unexplainable decision-making processes, and the

perpetuation of existing societal disparities through biased datasets and algorithms. Each of these risks significantly impacts patient safety, the trustworthiness of AI-enabled applications, and the ethical integrity of care delivery. To address these challenges, this essay advocates for the implementation of a multilayered governance model that collectively ensures a balanced and holistic governance approach.

References

Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V. I., & Precise4Q Consortium. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making*, *20*, 1-9.

Bak, M., Madai, V. I., Fritzsche, M. C., Mayrhofer, M. T., & McLennan, S. (2022). You can't have AI both ways: balancing health data privacy and access fairly. *Frontiers in Genetics*, *13*, 1490.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Broadhead, G. (2023). A Brief Guide to LLM Numbers: Parameter Count vs. Training Size. *Medium*. https://medium.com/@greg.broadhead/a-brief-guide-to-llm-numbers-parameter-count-vs-training-size-894a81c9258

Carlini, Nicholas, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts et al. "Extracting training data from large language models." In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633-2650. 2021.

Center for Research on Foundation Models. (n.d.). Home. *Stanford University*. https://crfm.stanford.edu/

Cheng, Y., Wang, F., Zhang, P., & Hu, J. (2016, June). Risk prediction with electronic health records: A deep learning approach. In *Proceedings of the 2016 SIAM international conference on data mining* (pp. 432-440). Society for Industrial and Applied Mathematics.

Chu, L. C., Anandkumar, A., Shin, H. C., & Fishman, E. K. (2020). The potential dangers of artificial intelligence for radiology and radiologists. *Journal of the American College of Radiology*, *17*(10), 1309-1311.

Cohen, I. G., & Mello, M. M. (2019). Big data, big tech, and protecting patient privacy. *Jama*, *322*(12), 1141-1142.

Cox, J. (2023). Facebooks' powerful large language model leaks online. *Vice.* Retrieved from https://www.vice.com/en/article/xgwqgw/facebooks-powerful-large-language-model-leaks-online-4chan-llama

Dinnerstein v. Google and The University of Chicago Medical Center 1:19cv—04311 (N.D. Ill.). (2019). https://www.courtlistener.com/docket/15841645/dinerstein-v-google-llc/#entry-1

Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., ... & Sui, Z. (2022). A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Elsayed, G. F., Goodfellow, I., & Sohl-Dickstein, J. (2018). Adversarial reprogramming of neural networks. *arXiv preprint arXiv:1806.11146*.

European Union. (2022). Artificial Intelligence Act. *EUR-Lex*. https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF

European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679

Fakoor, R., Ladhak, F., Nazi, A., & Huber, M. (2013, June). Using deep learning to enhance cancer diagnosis and classification. In *Proceedings of the international conference on machine learning* (Vol. 28, pp. 3937-3949). New York, NY, USA: ACM.

Floridi, L., & Cowls, J. (2022). A unified framework of five principles for AI in society. *Machine learning and the city: Applications in architecture and urban design*, 535-545.

Gaffney, T. (2023). Synthetic data generation: Building trust by ensuring privacy and quality. *IBM.* https://www.ibm.com/blog/synthetic-data-generation-building-trust-by-ensuring-privacy-and-quality/

Gymrek, M., McGuire, A. L., Golan, D., Halperin, E., & Erlich, Y. (2013). Identifying personal genomes by surname inference. *Science*, *339*(6117), 321-324.

Future of Life Institute. (2023). An open letter calling for a pause on all giant AI experiments. *Future of Life Institute*. https://futureoflife.org/open-letter/pause-giant-ai-experiments/

Futurescan. (2023). Consumer trends. *Futurescan: Healthacre Trends and Implications.*

Keddell, E. (2019). Algorithmic justice in child protection: Statistical fairness, social justice and the implications for practice. *Social Sciences*, *8*(10), 281.

Kiser, S., & Maniam, B. (2021). Ransomware: Healthcare industry at risk. *Journal of Business and Accounting*, *14*(1), 64-81.

Kleinman, Z., & Vallance, C. (2023). AI 'godfather' Geoffrey Hinton warns of dangers as he quits Google. *BBC News*. https://www.bbc.com/news/world-us-canada-65452940

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

Liddell, K., Simon, D. A., & Lucassen, A. (2021). Patient data ownership: who owns your health?. *Journal of Law and the Biosciences*, *8*(2), lsab023.

Liu, C. F., Chen, Z. C., Kuo, S. C., & Lin, T. C. (2022). Does AI explainability affect physicians' intention to use AI?. *International Journal of Medical Informatics*, *168*, 104884.

Makridakis, S. (2017). The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. *Futures*, *90*, 46-60.

Mäntymäki, M., Minkkinen, M., Birkstedt, T., & Viljanen, M. (2022). Putting AI ethics into practice: the hourglass model of organizational AI Governance. *arXiv preprint arXiv:2206.00335*.

Miotto, R., Wang, F., Wang, S., Jiang, X., & Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, *19*(6), 1236-1246.

Na, L., Yang, C., Lo, C. C., Zhao, F., Fukuoka, Y., & Aswani, A. (2018). Feasibility of reidentifying individuals in large national physical activity data sets from which protected health information has been removed with use of machine learning. *JAMA network open*, *1*(8), e186040-e186040.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447-453.

OpenAI. (2018). AI and Compute. *OpenAI*. https://openai.com/research/ai-and-compute

Power, H. M. L. C. C., & Progress, D. A. I. AI and Compute.

Rani, V., Nabi, S. T., Kumar, M., Mittal, A., & Kumar, K. (2023). Self-supervised learning: A succinct review. *Archives of Computational Methods in Engineering*, *30*(4), 2761-2775.

Rudin, C. (2018). Please stop explaining black box models for high stakes decisions. *Stat*, *1050*, 26.

Song, C., & Shmatikov, V. (2019). Overlearning reveals sensitive attributes. *arXiv preprint arXiv:1905.11742*.

Standards for Privacy of Individually Identifiable Health Information, 45 C.F.R. Part 164. (2000). https://www.federalregister.gov/documents/2000/12/28/00-32678/standards-for-privacy-of-individually-identifiable-health-information#sectno-reference-164.102

The Allan Turing Institute. (2023). Exploring foundation models - Session 1 [Video]. *YouTube*. https://www.youtube.com/watch?v=n9OkJBluOa4&t=3259s

Tuffaha, M. (2023). The Impact of Artificial Intelligence Bias on Human Resource Management Functions: Systematic Literature Review and Future Research Directions. *European Journal of Business and Innovation Research*, *11*(4), 35-58.

Van Dijck, G. (2022). Predicting recidivism risk meets AI Act. *European Journal on Criminal Policy and Research*, *28*(3), 407-423.

White House. (2022). Notice and Explanation. *AI Bill of Rights*. https://www.whitehouse.gov/ostp/ai-bill-of-rights/notice-and-explanation/

World Economic Forum. (2016). *WEF Values and the Fourth Industrial Revolution White Paper*. https://www3.weforum.org/docs/WEF_Values_and_the_Fourth_Industrial_Revolution_WHITEPAPER.pdf