Foundation Models in Healthcare: Opportunities, Risks & Strategies Forward

Anja Thieme
anthie@microsoft.com
Microsoft Health Futures
UK

Aditya Nori
adityan@microsoft.com
Microsoft Health Futures
UK

Marzyeh Ghassemi
mghassem@mit.edu
Massachusetts Institute of Technology
USA

Rishi Bommasani
nlprishi@stanford.edu
Stanford University
USA

Tariq Osman Andersen
tariq@di.ku.dk
University of Copenhagen
DK

Ewa Luger
ewa.luger@ed.ac.uk
The University of Edinburgh
UK

ABSTRACT

Foundation models (FMs) are a new paradigm in AI. First pretrained on broad data at *immense* scale and subsequently adapted to more specific tasks, they achieve high performances and unlock powerful new capabilities to be leveraged in many domains, including healthcare. This SIG will bring together researchers and practitioners within the CHI community interested in such emerging technology and healthcare. Drawing attention to the rapid evolution of these models and proposals for their wide-spread adoption, we aim to demonstrate their strengths whilst simultaneously highlighting deficiencies and limitations that give raise to ethical and societal concerns. In particular, we will invite the community to actively debate how the field of HCI – with its research frameworks and methods – can help address some of these existing challenges and mitigate risks to ensure the safe and ethical use of the end-product; a requirement to realize many of the ambitious visions for how these models can positively transform healthcare delivery. This conversation will benefit from a diversity of voices, critical perspectives, and open debate, which are necessary to bring about the right norms and best practices, and to identify a path forward in devising responsible approaches to future FM design and use in healthcare.

CCS CONCEPTS

• **Human-centered computing → HCI design and evaluation methods**; • **Computing methodologies → Artificial intelligence**.

KEYWORDS

Foundation models, healthcare, responsible AI, ethics, socio-technical systems, interaction design

ACM Reference Format:
Anja Thieme, Aditya Nori, Marzyeh Ghassemi, Rishi Bommasani, Tariq Osman Andersen, and Ewa Luger. 2023. Foundation Models in Healthcare: Opportunities, Risks & Strategies Forward. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3544549.3583177

CHI EA '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3583177

1 INTRODUCTION

The latest wave of AI innovation sees the evolution of a new class of AI models often referred to as foundation models (FMs) — a term popularized by the Stanford Institute for Human-Centered AI [5]. Recent examples include models like Google's LaMDA [10] and OpenAI's GPT-3.5 [28] that demonstrate impressive capabilities to generate coherent text; or OpenAI's DALL·E 2 [33], which can create realistic images and art from a text description. These models are based on deep learning and trained mostly via self-supervision on broad data at *immense scale* and high resource costs. The resulting general-purpose models are powerful and complex – often containing billions, even trillions of model parameters – and can be adapted to a wide range of downstream tasks [5].

1.1 Paradigm Shift in AI

The *pretraining then task-adaptation* approach to FMs presents a paradigm shift away from a more traditional focus on task-specific models that have dominated the AI landscape thus far. Instead, pretrained models can be re-used and adapted to specific tasks for which they were not specifically trained [5, 39]. Task adaptation can be achieved in multiple ways, e.g., via user or engineer prompts; continual learning; or a process of fine-tuning - whereby incremental adjustments to the model are typically learned from a much smaller training sample [38] as would be required by more traditional ML approaches. This expands opportunities for domains in which insufficient data is an obstacle for training task-specific algorithms [39], and invites considerable excitement about a future of more flexible, re-usable AI models that can be scaled and applied to any domain or task. This includes healthcare, where FMs may offer potentially remarkable new technology capabilities that could completely transform clinical practice [40].

1.2 Opportunities for Healthcare

Existing proposals for healthcare applications include the adaptation of FMs to achieve increased efficiency in specific tasks related to diagnosis and treatment (e.g., disease prediction [34], triage or

discharge recommendations [21]), and assistance with health administration tasks [5] via workflow optimizations (e.g., clinical notes summarization [22] or medical text simplication [19]).

FMs also present as a central storage of medical knowledge that could be queried by healthcare professionals or the public, suggesting its use in medical question-answering [26] and chatbot applications [12]. For example, recent developments like ChatGPT [28], Bing Chat [24], or Bard [30] offer conversational UIs to help people identify and extract useful insights and deepen understanding of information, which can become adopted to search for health advice.

Furthermore, new FM-enabled services may not only accelerate healthcare application development, but also research by providing capabilities to automate, for example: structured data set generation [32]; data labelling; or aid synthetic data creation [7]. Future work may also explore the development of entirely new FM-enabled capabilities that could be afforded especially through multi-modal data that the healthcare domain is particularly characteristic for. Extending beyond natural language processing, we already find examples in biomedical research that demonstrate great advances in predicting human proteins (e.g., AlphaFold [37]) to assist structure-based drug development; alongside efforts in genome sequencing to speed-up detection of variants that cause genetic disease [17], through to proposals for optimized clinical trial design [8].

1.3 What makes FMs Powerful makes them Risky: Emergence & Homogeneity

To gain their power, foundation models leverage deep learning (DL) approaches, whereby higher-level features *automatically emerge* from the raw data inputs as an implicitly induced learning process [5] – based on values the model itself chooses and optimizes during training. A central feature of deep learning is the possibility to pretrain the model on large volumes of unannotated datasets via *self-supervised* processes, which – contrary to earlier generations of AI systems that relied on the curation of medical knowledge by experts and required explicit expressions of robust decision rules and labelled input-output pairs [40] – does not require any human input. Self-supervised learning therefore is what unlocks an entire new scale of data analysis, resulting in performance gains [5]; all of which is powered by advances in GPU throughput, memory capacity, and model architectures like transformer networks [2, 39].

Subsequent pretrained models are then adapted to other tasks using *transfer learning* [34], whereby knowledge inferred from the intrinsic structure of data in one task becomes applied to another [39]. In other words, the powerful FMs that evolve, for example Google's language model BERT [13], then come to serve as the basis for new state-of-the art models as adaptations from that FM (e.g., CXR-BERT [3], Med-BERT [34] or PubMedBERT [18] in healthcare). However, this *homogenization*, whereby few models become repeatedly re-used as basis for other applications to increases cost-effectiveness of AI systems, is simultaneously a key concern since any inherent defects of the FM become inherited by all models fine-tuned on them [4, 5].

Furthermore, the scale, complexity and emergent nature of their learning process make it difficult, if not impossible, to understand how FMs and their derivatives work, or when they might fail [5]. Given early stages of FM development that leaves many of their potential pitfalls under-explored, this requires particular caution for any prospective use within sensitive, high-stakes domains like healthcare to not accentuate risks of harm (i.e., how the use of FM-based applications may exacerbate social inequalities) [5, 39]; and raises fundamental questions about the responsible, ethical and safe use of such technologies going forward [2].

2 FM USE IN HEALTHCARE: CHALLENGES

Next, we extend on some of these risks of FM use and what reinforces them; suggesting topics for conversation to invite active debate at CHI and solicit proposals for how the research community may be able to help with addressing these issues.

2.1 Data is not Neutral & Algorithms are not Objective: Health Disparities & Societal Bias

Data is the building block of sense-making in AI [11]. In healthcare, data can be widely sourced (e.g., from care providers, insurers, publications) [39], and vary in type (e.g., clinical notes, medical images), scale (e.g., patient vs. population level), or style (professional vs. lay language) [5]; suggesting unique opportunities for multi-modal FMs as well as core challenges for their training (e.g., across patient cohorts; bias exacerbation). While ML methods are generally considered well placed to handle and derive useful insights from large volumes of diverse, multi-dimensional data [35], *it is important to recognize how societal bias and inequality manifest in data*. Disparities range from the types of healthcare problems that are being prioritized and funded (e.g., as downstream applications of FMs); through to the exclusion of specific population groups and their misrepresentation in data collected. This can be due to lack of access to healthcare; strict restrictions in patient criteria for participation in clinical research trials; or higher risks of inaccurate data capture due to documentation errors and systemic discrimination of individuals from disadvantaged communities [6]. For example, having to verify citizenship at hospitals in California meant that the rate of autism diagnosis for Hispanic children, who are often undocumented immigrants, fell following aggressive federal anti-immigration policies [15].

Aside from bias and harm resulting from: underrepresentation, overrepresentation or misrepresentation of certain groups in datasets that serve as the training ground for FMs and their derivatives [39], Bender et al. [2] discuss how stereotypical and derogatory associations along gender, race, ethnicity, and disability status are encoded in large language models (LLMs) that are often built on uncurated, static datasets mostly crawled from the internet (for GPT) or Wikipedia and online books (for BERT), which entrenches dominant viewpoints and reinforces inequalities (e.g., the term 'women doctors' subtly implies the term doctor entails not-women and excludes non-binary gender identities). Generating text that amplifies underlying bias leads to "harms of subjugation, denigration, belittlement, loss of opportunity and others on the part of those discriminated against" [ibid]. With *societal bias being deeply rooted within language*, post-hoc filtering mechanisms alone that reduce occurrences of unintelligible or bad content in training data are insufficient, nor will larger datasets be a guarantor for greater diversity. This suggests a shift towards more thoughtful dataset curation [2, 5], alongside efforts to increase systematic reporting

and risk analysis (e.g., via datasheets, impact assessments); and to provide appropriate training resources to sensitize towards these broader social and ethical issues.

Furthermore, *in subsequent algorithm design, model developers make important choices* in what they optimize model performance for [35]; to what extent specific sensitive attributes are accounted for in model design and testing; and how desired outcomes are defined and measured [6], which can further exacerbate disparities. For example, an algorithm used to optimize referrals to long-term care programs gave a similar risk score to Black patients as White patients, when Black patients were considerably sicker and needed extra care – a racial biases due to care costs being used as proxy for health by some measure of predictive accuracy; yet unequal care access can mean less money is spend caring for Black patients [27].

An added difficulty in the FM pretraining fine-tuning paradigm is to understand where in the model ecosystem harms occur, and how responsibilities for initial model development and subsequent downstream fine-tuning become allocated [5, 39]. Alongside calls for appropriate bias detection techniques, it is important that development teams do not approach challenges of societal disparities and bias solely as a technical problem that can be 'engineered out', but seek to better understand, i.e.,: how sensitive attributes and other confounding factors relate to the outcome of interest and can be causes of downstream harm [6]; or how different types of errors may disproportionately affect different patient groups or health service providers [36]. Recent research by Adam et al. [1] also showed how framing the AI output in communications to end-users can help mitigate discriminatory effects of biased AI advice; raising promise in careful interaction design.

2.2 Over-trusting High Performance & Output Coherence: Ensuring Safe & Reliable Use

Whilst it is evident that larger models achieve higher accuracy and unlock new technology capabilities, it is important to also bring careful consideration to ethical and legal requirements for their use to be safe, fair, to protect peoples' privacy [39], and ultimately, to benefit patients and care providers. Safe use in healthcare implies that the system provides factually accurate, reliable information for clinical decision-making. However, being able to assess whether FM-derived application outputs are correct or not becomes increasingly difficult, even for experts. This is exemplified by ChatGPT, whose coherent text generation feels indistinguishable from language produced by humans [21], leading us to interpret the generated text as meaningful and truthful, which in turn increases risk of automation bias and opportunities for deliberate misuses (e.g., bad actors creating false, manipulative contents) [2]. And yet, *even highly plausible-sounding outputs can be incorrect.* To better manage risks resulting from AI errors, the field of explainable AI (XAI) has brought forward various techniques to explain model workings to facilitate scrutiny and enable contestation of offered results. However, in practice, clinicians often lack the extra time and mental capacity required to engage with such explanations, which also assume technical expertise and clinician interest in wanting to interrogate AI outputs [36]. Whilst most post-hoc explanations can be useful at an aggregate level (e.g., for model improvements or audit), they are often unreliable and less useful for individual instances that matter for justifying person-focused healthcare decisions [16, 20]. Research has also shown that the addition of explanations can in fact increase rather than decrease over-reliance on AI outputs (e.g., due to anchoring and confirmation bias) [14, 31]. That aside, given the scale and complexity of FMs and their derivatives, it may be too difficult, if not impossible, to understand their workings. Amongst others, this suggests more careful consideration: in choices when and when not to deploy FM-enabled applications in healthcare [21]; for training and interface design (on-boarding and use) to make transparent the limitations and probabilistic nature of AI outputs; and to shift the focus more towards the development of rigorous and thorough validation procedures [16]. Here we might ask: what types of internal and external validations and continued testing and monitoring approaches would be needed as guarantors that these models are safe and reliable when in use; serve their intended purpose(s); and do not unfairly discriminate against specific person characteristics or population groups?

2.3 Building AI in a Vacuum: Decontextualized & Centralized

In what has been described as "a race for getting the technology right before exposing human-end users to new promising AI tools" [29]; the tendency to develop AI "in a vacuum" [25] – disconnected from well-defined needs of intended uses and downstream use contexts, is increasingly criticized [23]. The *decontextualized* treatment of AI development has led to more calls for ethnographic studies to better understand actual practices, applications and uses that surround algorithmic technologies within their use environments. This is particularly complicated in the context of FMs, which – as the term "foundation" suggests – present an early component within the AI system development pipeline that needs further adaptation to be useful [5]. Nonetheless, developing a better understanding of especially the limitations of FM-enabled applications in context can help: to reduce misleading hypes about their capabilities and instead ensure a closer focus on the needs and use context of the people who are intended to benefit from these technologies as well as those, who may be adversely affected [2]; to identify context-appropriate risk mitigation strategies (e.g., learning from human-peer review practices in healthcare to identify and address errors); or to encourage new research directions that do not necessarily depend on having FMs [ibid]. It is also essential to close current gaps in moving from compelling technical proof-of-concepts and successful lab experiments towards the integration and deployment of AI-enabled systems within routine care [36]. To date, only few experimental studies examined whether AI models achieve their intended effects when deployed in the real-world [29, 31]. However, to validate clinical utility, and enable for new AI systems to reach their full potential, requires their end-to-end design, integration and continued performance monitoring within clinical workflows. While the challenge of translating AI research into successfully deployable clinical applications is not straight forward [40], the HCI community is well equipped with frameworks and methods to drive forward what has been termed the "last mile" of AI in healthcare [9, 29]. In this regard, Zajac et al. [41] provide detailed insights and concrete guidance to HCI researchers and practitioners about requirements for effectively realizing ML in medical practice.

This journey has the added complication that broader inter-disciplinary engagement and contributions to FM research is constrained to mostly a small number of high-resourced industry labs or large non-profits like OpenAI, who have the financial means and technology infrastructure to train FMs in the first place. Such *centralization* of power excludes much of the larger academic research community, thereby reducing diversity in perspectives on how FM development should be shaped [5, 11]. How to encourage more inter-disciplinary engagement and discourse? And what should and could industries and governments do to enable more resource sharing and equitable access to compute infrastructures?

REFERENCES

[1] Hammaad Adam, Aparna Balagopalan, Emily Alsentzer, Fotini Christia, and Marzyeh Ghassemi. 2022. Mitigating the impact of biased artificial intelligence in emergency decision-making. *Communications Medicine* 2, 1 (2022), 1–6.

[2] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 610–623.

[3] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, et al. 2022. Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing. *arXiv preprint arXiv:2204.09817* (2022).

[4] Rishi Bommasani, Kathleen Creel, Ananya Kumar, Dan Jurafsky, and Percy Liang. 2022. Picking on the Same Person: Does Algorithmic Monoculture lead to Outcome Homogenization?. In *Advances in Neural Information Processing Systems*, Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (Eds.). https://openreview.net/forum?id=-H6kKm4DVo

[5] Rishi Bommasani, Drew A Hudson, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[6] Irene Y Chen, Emma Pierson, Sherri Rose, Shalmali Joshi, Kadija Ferryman, and Marzyeh Ghassemi. 2021. Ethical machine learning in healthcare. *Annual review of biomedical data science* 4 (2021), 123–144.

[7] Richard J Chen, Ming Y Lu, Tiffany Y Chen, Drew FK Williamson, and Faisal Mahmood. 2021. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering* 5, 6 (2021), 493–497.

[8] Isabel Chien, Nina Deliu, Richard Turner, Adrian Weller, Sofia Villar, and Niki Kilbertus. 2022. Multi-disciplinary fairness considerations in machine learning for clinical trials. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 906–924.

[9] Enrico Coiera. 2019. The last mile: where artificial intelligence meets reality. *Journal of medical Internet research* 21, 11 (2019), e16323.

[10] Eli Collins and Zoubin Ghahramani. 2021. LaMDA: our breakthrough conversation technology. https://blog.google/technology/ai/lamda/

[11] Kate Crawford. 2021. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press, New Haven and London.

[12] Dina Demner-Fushman, Yassine Mrabet, and Asma Ben Abacha. 2020. Consumer health information and question answering: helping consumers find answers to their health-related information needs. *Journal of the American Medical Informatics Association* 27, 2 (2020), 194–201.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[14] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The impact of placebic explanations on trust in intelligent systems. In *Extended abstracts of the 2019 CHI conference on human factors in computing systems*. 1–6.

[15] Christine Fountain and Peter Bearman. 2011. Risk as social context: immigration policy and autism in California 1. In *Sociological Forum*, Vol. 26. Wiley Online Library, 215–240.

[16] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. 2021. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health* 3, 11 (2021), e745–e750.

[17] Sneha D Goenka, John E Gorzynski, Kishwar Shafin, Dianna G Fisk, Trevor Pesout, Tanner D Jensen, Jean Monlong, Pi-Chuan Chang, Gunjan Baid, Jonathan A Bernstein, et al. 2022. Accelerated identification of disease-causing variants with ultra-rapid nanopore genome sequencing. *Nature Biotechnology* 40, 7 (2022), 1035–1041.

[18] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, et al. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)* 3, 1 (2021), 1–23.

[19] Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Sabel, Jens Ricke, et al. 2022. ChatGPT Makes Medicine Easy to Swallow: An Exploratory Case Study on Simplified Radiology Reports. *arXiv preprint arXiv:2212.14882* (2022).

[20] Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*. PMLR, 1885–1894.

[21] Diane M Korngiebel and Sean D Mooney. 2021. Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 (GPT-3) in healthcare delivery. *NPJ Digital Medicine* 4, 1 (2021), 1–3.

[22] Kundan Krishna, Sopan Khosla, Jeffrey P Bigham, and Zachary C Lipton. 2020. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. *arXiv preprint arXiv:2005.01795* (2020).

[23] Q Vera Liao, Yunfeng Zhang, Ronny Luss, et al. 2022. Connecting Algorithmic Research and Usage Contexts: A Perspective of Contextualized Evaluation for Explainable AI. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 10. 147–159.

[24] Yusuf Mehdi. 2023. Reinventing search with a new AI-powered Microsoft Bing and Edge, your copilot for the web. https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/

[25] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547* (2017).

[26] Emmanuel Mutabazi, Jianjun Ni, Guangyi Tang, and Weidong Cao. 2021. A review on medical textual question answering systems based on deep learning approaches. *Applied Sciences* 11, 12 (2021), 5456.

[27] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[28] OpenAI. 2022. CHATGPT: Optimizing language models for dialogue. https://openai.com/blog/chatgpt/

[29] Tariq Osman Andersen, Francisco Nunes, Lauren Wilcox, Elizabeth Kaziunas, Stina Matthiesen, and Farah Magrabi. 2021. Realizing AI in healthcare: challenges appearing in the wild. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–5.

[30] Sundar Pichai. 2023. An important next step on our AI journey. https://blog.google/technology/ai/bard-google-ai-search-updates/

[31] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–52.

[32] Sam Preston, Mu Wei, Rajesh Rao, Robert Tinn, Naoto Usuyama, Michael Lucas, Roshanthi Weerasinghe, Soohee Lee, Brian Piening, Paul Tittel, et al. 2022. Towards Structuring Real-World Data at Scale: Deep Learning for Extracting Key Oncology Information from Clinical Text with Patient-Level Supervision. *arXiv preprint arXiv:2203.10442* (2022).

[33] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* (2022).

[34] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine* 4, 1 (2021), 1–13.

[35] Luc Rubinger, Aaron Gazendam, Seper Ekhtiari, and Mohit Bhandari. 2022. Machine learning and artificial intelligence in research and healthcare. *Injury* (2022).

[36] Anja Thieme, Maryann Hanratty, Maria Lyons, Jorge E Palacios, Rita Marques, et al. 2022. Designing Human-Centered AI for Mental Health: Developing Clinically Relevant Applications for Online CBT Treatment. *ACM Transactions on Computer-Human Interaction* (2022).

[37] Kathryn Tunyasuvunakool, Jonas Adler, Zachary Wu, Tim Green, Michal Zielinski, Augustin Žídek, Alex Bridgland, Andrew Cowie, Clemens Meyer, Agata Laydon, et al. 2021. Highly accurate protein structure prediction for the human proteome. *Nature* 596, 7873 (2021), 590–596.

[38] Walter F Wiggins and Ali S Tejani. 2022. On the Opportunities and Risks of Foundation Models for Natural Language Processing in Radiology. *Radiology: Artificial Intelligence* 4, 4 (2022), e220119.

[39] Malwina Anna Wójcik. 2022. Foundation Models in Healthcare: Opportunities, Biases and Regulatory Prospects in Europe. In *International Conference on Electronic Government and the Information Systems Perspective*. Springer, 32–46.

[40] Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane. 2018. Artificial intelligence in healthcare. *Nature biomedical engineering* 2, 10 (2018), 719–731.

[41] Hubert D Zajac, Dana Li, Xiang Dai, Jonathan F Carlsen, Finn Kensing, and Tariq O Andersen. 2023. Clinician-facing AI in the Wild: Taking Stock of the Sociotechnical Challenges and Opportunities for HCI. *ACM Transactions on Computer-Human Interaction* (2023).