Volume Title: The Economics of Artificial Intelligence: Health Care Challenges

Volume Authors/Editors: Ajay Agrawal, Joshua Gans, Avi Goldfarb, and Catherine Tucker, editors

Volume Publisher: University of Chicago Press

Volume ISBNs: 978-0-226-83311-8 (cloth); 978-0-226-83312-5 (electronic)

Volume URL: https://www.nber.org/books-and-chapters/economics-artificial-intelligence-health-care-challenges

Conference Date: September 22–23, 2022

Publication Date: March 2024

Chapter Title: Comment on "Health Data Platforms”

Chapter Author(s): Tyna Eloundou, Pamela Mishkin

Chapter URL: https://www.nber.org/books-and-chapters/economics-artificial-intelligence-health-care-challenges/comment-health-data-platforms-eloundou

Chapter pages in book: p. 95 – 97

References

Abelson, R., and J. Creswell. 2012. "Hospital Chain Inquiry Cited Unnecessary Cardiac Work." *New York Times*, August 6, 2012.

Agrawal, A., J. Gans, and A. Goldfarb. 2018. *Prediction Machines: The Simple Economics of Artificial Intelligence*. Boston, MA: Harvard Business Press.

Currie, J., and W. B. MacLeod. 2017. "Diagnosing Expertise: Human Capital, Decision Making, and Performance among Physicians." *Journal of Labor Economics* 35 (1): 18977. https://doi.org/10.1086/687848.

Donoho, D. 2017. "50 Years of Data Science." *Journal of Computational and Graphical Statistics* 26 (4): 745–66. https://doi.org/10.1080/10618600.2017.1384734.

Hill, R., C. Stein, and H. Williams. 2020. "Internalizing Externalities: Designing Effective Data Policies." *AEA Papers and Proceedings* 110: 49–54. https://doi.org/10.1257/pandp.20201060.

Kaushal, A., R. Altman, and C. Langlotz. 2020. "Geographic Distribution of US Cohorts Used to Train Deep Learning Algorithms." *JAMA* 324 (12): 1212–13. https://doi.org/10.1001/jama.2020.12067.

Kleinberg, J., J. Ludwig, S. Mullainathan, and Z. Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review* 105 (5): 491–95. https://doi.org/10.1257/aer.p20151023.

Mullainathan, S., and Z. Obermeyer. 2022a. "Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care." *Quarterly Journal of Economics* 137 (2): 1–51. https://doi.org/10.1093/qje/qjab046.

Mullainathan, S., and Z. Obermeyer. 2022b. "Solving Medicine's Data Bottleneck: Nightingale Open Science." *Nature Medicine* 28: 897–99. https://doi.org/10.1038/s41591-022-01804-4.

Neumann, P. J., J. T. Cohen, and M. C. Weinstein. 2014. "Updating Cost-Effectiveness—The Curious Resilience of the $50,000-per-QALY Threshold." *New England Journal of Medicine* 371: 796–97. https://doi.org/10.1056/NEJMp1405158.

Price, W. N., and I. G. Cohen. 2019. "Privacy in the Age of Medical Big Data." *Nature Medicine* 25: 37–43. https://doi.org/10.1038/s41591-018-0272-7.

Stern, A. D. 2022. "The Regulation of Medical AI: Policy Approaches, Data, and Innovation Incentives." NBER Working Paper No. 30639. Cambridge, MA: National Bureau of Economic Research. https://doi.org/10.3386/w30639.

Comment Tyna Eloundou and Pamela Mishkin

Machine learning tools like large language models (LLMs) have shown remarkable improvements in capabilities in various natural language processing tasks, such as text generation, summarization, and dialogue, over the last few years. However, developing and deploying these models in a responsible and beneficial manner requires access to high-quality and diverse datasets that reflect the domains and contexts of interest. In the field of language

Tyna Eloundou and Pamela Mishkin are researchers at OpenAI.

For acknowledgments, sources of research support, and disclosure of the authors' material financial relationships, if any, please see https://www.nber.org/books-and-chapters/economics-artificial-intelligence-health-care-challenges/comment-health-data-platforms-eloundou.

modeling, recent advances have highlighted the importance of using small, curated datasets to fine-tune LLMs that have been pretrained on massive amounts of unstructured web data (Ouyang et al. 2022). In this comment, we discuss two ways that academic researchers can contribute to the responsible development and deployment of useful machine learning models: creating and governing access to structured datasets (as exemplified by Dandelion and Nightingale) and examining the impacts and policy implications of these systems.

OpenAI is an organization committed to ensuring that artificial general intelligence benefits all of humanity (OpenAI 2018). To this end, we are focused on building and deploying large language models such as GPT-3, Codex, and DALL-E responsibly. To do so, we must carefully track how they are used—and misused—to inform our future research and deployment decisions (Brundage et al. 2022).

One of the key challenges we face is gathering the vast amounts of data needed to train and fine-tune these models. Our current approach consists of two steps: we first train a baseline model with unstructured web data while taking care to ensure a minimum level of quality, and we then fine-tune the models using structured, high-quality data that has been annotated by human contractors. This two-pronged approach is essential for developing models that both have general language understanding and local context (instructions, examples, or prompts) about the use-cases that are presented to them. You can learn more about such techniques via projects like InstructGPT (Ouyang et al. 2022), where a language model that has language understanding is trained to better respond to human instructions, and Sparrow (Glaese et al. 2022), where an information-seeking dialogue agent is made to behave according to particular rules and requirements. This is similar to how doctors learn in school: they first receive general education before specializing in particular domains.

The second step of fine-tuning models is the most difficult and demanding, as it requires high-quality and detailed data in domains that often require some level of expertise and domain knowledge. Moreover, creating and using such data in a responsible manner entails addressing several ethical and social issues, such as ensuring fair compensation, informed consent, and representativeness of the data contributors and the potential beneficiaries and stakeholders of the technology. The healthcare field is one of the domains where such issues are particularly salient and challenging, and where there is a significant gap between the availability and quality of data and the needs and expectations of the AI development community.

We believe that initiatives like Dandelion and Nightingale can play a vital role in bridging this gap and facilitating the responsible development and deployment of LLMs in the healthcare domain. Dandelion is a project that partners with large health systems to create and provide access to world-class, high-quality medical datasets that cover various aspects of healthcare, such as diagnosis, treatment, and outcomes. Nightingale is a project that partners with researchers to grant them access to large and diverse datasets and compute resources that can enable them to conduct novel and impactful research in various subfields of healthcare, such as epidemiology, genomics, and clinical trials. We're pleased to see these projects adopting best practices like on data governance and data documentation (Gebru et al. 2021).

We are excited about the potential of these projects to improve the state of the art and the state of the practice of LLMs in the healthcare domain. However, we also recognize that there are still many open and important questions about the impacts and implications of using these systems in real-world settings, such as their effects on the quality and equity of healthcare delivery, their risks and challenges for privacy and security, and their ethical and legal ramifications. We encourage both projects to also provide datasets that can help researchers and practitioners to address these questions, to better understand the realized impacts when models are deployed in the real world.

References

Brundage, Miles, Katie Mayer, Tyna Eloundou, Sandhini Agarwal, Steven Adler, Gretchen Krueger, Jan Leike, and Pamela Mishkin. (2022). "Lessons Learned on Language Model Safety and Misuse." OpenAI. https://openai.com/blog/language-model-safety-and-misuse/.

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. "Datasheets for Datasets." *Communications of the ACM* 64 (12): 86–92.

Glaese, Amelia, Nat McAleese, Maja Trębacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, et al. 2022. "Improving Alignment of Dialogue Agents via Targeted Human Judgements." arXiv preprint. arXiv: 2209.14375.

OpenAI. 2018. "OpenAI Charter." https://openai.com/charter/

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. "Training Language Models to Follow Instructions with Human Feedback." arXiv preprint. arXiv: 2203.02155.

Comment Judy Gichoya

Introduction

In their chapter, Mullainathan and Obermeyer utilize a clinical use case for predicting sudden cardiac death from electrocardiograms (ECGs) to

Judy Gichoya is an associate professor in the Department of Radiology and Imaging Sciences at Emory University School of Medicine.

For acknowledgments, sources of research support, and disclosure of the author's material financial relationships, if any, please see https://www.nber.org/books-and-chapters/economics-artificial-intelligence-health-care-challenges/comment-health-data-platforms-gichoya.