From Data Hacking to Data Excellence: The Role of HCI in High-stakes AI

SHIVANI KAPANIA, Google Research, India
KRISTEN OLSON, Google Research, USA
HANNAH HIGHFILL, Google, USA
DIANA AKRONG, Google Research, Ghana
LORA AROYO, Google Research, USA
PRAVEEN PARITOSH, Google Research, USA
NITHYA SAMBASIVAN, Google Research, USA

AI models are increasingly built and deployed in high-stakes domains such as cancer diagnosis, poaching detection and predictive policing. Data is the critical infrastructure which largely determines necessary to build these AI systems. We studied how humans interact with data throughout the ML pipeline, and in this paper, we report on data practices in high-stakes AI, from a literature review on data labeling, and interviews with 53 AI practitioners in India, East and West African countries, and USA. We discuss how HCI can play a role in making high-stakes AI into a synergistic endeavour; being aware of, and asking the right questions of AI systems could help shift the focus from hacking to data excellence.

Additional Key Words and Phrases: datasets, data quality, artificial intelligence, high-stakes AI, data collectors, incentives

1 INTRODUCTION

One of the biggest determinants of AI success is availability of good data [4]. However, bad data are very common, leading to 51% of models stalling from production. Why are bad data so prevalent, despite the numerous data wrangling and analysis tools, open & commercial data sets, and labelling platforms? Why is data quality adoption so poor? What are developers prioritising over good data? What causes poor quality in data? Research shows that data quality improvements are as valuable as 'better' algorithms [3].

In order to answer these questions, we attempted to understand the various human factors and practices that influence the decisions that go into dataset quality, coverage, and freshness. Humans are involved in every step of the AI data pipeline, from labelling to model tuning. We need a shift in thinking from merely data cleaning support to a better understanding of how and why the *humans in the AI pipeline* (from data collectors, annotators to AI practitioners) work with data, and improved metrics of developer agency, incentives, and performance. To understand this problem, we focused on HCI research on data practices and challenges among AI developers and AI researchers, *within high-stakes domains*. Human Computer Interaction (HCI) brings forth a systemic perspective on designing real world AI systems, such as identifying real human needs, mapping them onto AI models, designing the right datasets and algorithms, building usable and useful interfaces, adding safeguards, and evaluating with real users before scaling. Our aim is to empower the humans in the AI pipeline, by using HCI techniques to understand how they interact with data and

Manuscript submitted to ACM

the challenges that they face. This includes the data collectors, data labellers, data scientists, AI practitioners and researchers, organizations and governments, and ultimately, users.

We believe that human-centered research and design can turn data acquisition into a clean, reliable step in the AI engineering process for high-stakes domains–resulting in better, safer, and robust systems for all. Tools and interfaces that used to work for traditional AI datasets, like Imagenet (very large-scale crowd work), may not work for high-stakes data like medical data, low-resource NLP data, or dialog data. Our research aims to create reproducible and repeatable assets of knowledge, user interface and process guidelines, and technical assets like open-source tools to improve the data acquisition process.

2 METHODOLOGY

As a part of our research project, we conducted interviews with AI practitioners (details below), a literature review on human computation, data labelling, and annotator well-being, and have interviews with data annotators ongoing. Between May and July 2020, we conducted semi-structured interviews with a total of **53** AI practitioners[1] working in high-stakes applications of AI development. Interviews were focused on (1) data sources and AI lifecycles; (2) defining data quality; (3) feedback loops from data quality; (4) upstream and downstream data effects; (5) stakeholders and accountability; (6) incentive structures; and (7) useful interventions. Each session focused on the participant's experiences, practices, and challenges in AI development and lasted about 75 minutes each. Participants signed informed consent documents acknowledging their awareness of the study purpose and researcher affiliation prior to the interview. In our sample, AI practitioners were located in, or worked primarily on projects based in, India **(23)**, the US **(16)**, or East and West African countries **(14)**. We sampled more widely in Africa due to the nascent AI Ecosystem compared to other continents [7], with 14 total interviews including Nigeria **(10)**, Kenya **(2)**, Uganda **(1)**, and Ghana **(1)**. We interviewed **45** male and **8** female AI practitioners. Refer to Sambasivan et al. [10] for complete methodological details of the interviews with AI practitioners.

3 FINDINGS

Acquiring usable and high quality data often feels like the messiest and least predictable part of the AI engineering process. Data is also the least glamorous and incentivized aspect, relative to building new models and algorithms. Our research shows that data quality is often sacrificed by AI developers due to competing factors like cost and time to complete projects, but this results in disastrous, compounding effects in the downstream on model accuracy, user trust, and business revenue. Data acquisition (collection, labelling, drifts) is very much an HCI problem of bringing visibility, control, trust, and explainability to underlying data, whether through new tools or from live data streams.

We define *Data Cascades* based on the empirical results in this study as *compounding events causing negative, downstream effects from data issues, that result in technical debt over time* [10]. In our study, **92%** practitioners experienced at least one cascade. Data cascades are influenced by, (a) the activities and interactions of actors involved in the AI development (developers, governments, and field partners), (b) the physical world and community in which the AI system is situated (hospitals in rural areas where sensor data collection occurs). Data cascades are complex, long-term, occur frequently and persistently; they are opaque in diagnosis and manifestation. In the absence of well-defined and timely signals, practitioners turned to proxy metrics (accuracy, precision, or F1 score), where the unit of measurement is the entire system, not datasets.

[1] Although our participants had different job roles (including, in research), all were focused on applied deployments in high-stakes domains.

Data cascades are triggered when conventional AI practices are applied in high-stakes domains, which are characterised by high accountability, inter-disciplinary work, and resource constraints. For example, practitioners viewed data as operations, moved fast, hacked model performance (through hyper-parameters rather than data quality), and did not appear to be fully equipped to recognise upstream and downstream issues and consequences. Data cascades have negative impacts, with some leading to harm to beneficiary communities, burnout of relationships with stakeholders, discarding entire datasets, and performing costly iterations. Cascades are often avoidable by step-wise and early interventions, which were exceptional in our study.

Thoughtful practices, though rare, such as good data documentation [9], training data collectors and annotators on aspects of quality, transparency about the purposes of data collection— all contributed towards better quality data and eventually avoiding data cascades. Our research also uncovered how designing usable interfaces for annotators [5, 12], conducting pilot and feedback sessions [1, 2], and understanding their perspectives contributed immense knowledge about the data and the task itself. To illustrate how an HCI approach can improve data quality and accuracy, take an example of how adding an 'I don't know' option to labelling interfaces led to an increase in accuracy of the labels–demonstrating how a design element can significantly impact data quality [11]. In another study, researchers show that simplified UIs and localized interfaces led to improved task completion among lower literate workers on rater platforms like Mechanical Turk [6]. We need innovation in interfaces, standards and processes for decentralized, specialized and high-stakes AI data. Conducting research with/for data collectors, raters, and AI developers will help us identify the right knobs for data quality, accuracy and fairness. We need to move from current approaches that are reactive and view data as 'grunt work' towards a proactive focus on *data excellence*—focusing on the practices, politics, and values of humans of the data pipeline to improve the quality and sanctity of data, through the use of processes, standards, infrastructure and incentives (and other interventions, as identified by Paritosh [8]).

REFERENCES

[1] Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. 2012. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*. 1013–1022.

[2] Snehalkumar Gaikwad, Nalin Chhibber, Vibhor Sehgal, Alipta Ballav, Catherine Mullings, Ahmed Nasser, Angela Richmond-Fuller, Aaron Gilbee, Dilrukshi Gamage, Mark Whiting, et al. 2017. Prototype tasks: improving crowdsourcing results through rapid, iterative task design. *arXiv preprint arXiv:1707.05645* (2017).

[3] Archana Ganapathi and Yanpei Chen. 2016. Data quality: Experiences and lessons from operationalizing big data. In *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, 1595–1602.

[4] Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems* 24, 2 (2009), 8–12.

[5] Ayush Jain, Akash Das Sarma, Aditya Parameswaran, and Jennifer Widom. 2017. Understanding workers, developing effective tasks, and enhancing marketplace dynamics: a study of a large crowdsourcing marketplace. *arXiv preprint arXiv:1701.06207* (2017).

[6] Shashank Khanna, Aishwarya Ratan, James Davis, and William Thies. 2010. Evaluating and improving the usability of Mechanical Turk for low-income workers in India. In *Proceedings of the first ACM symposium on computing for development*. 1–10.

[7] Hannah Miller and Richard Stirling. 2019. Government AI Readiness Index 2019 — Oxford Insights — Oxford Insights. https://www.oxfordinsights.com/ai-readiness2019. (Accessed on 09/14/2020).

[8] Praveen Paritosh, Matt Lease, Mike Schaekermann, and Lora Aroyo. 2020. First workshop on Data Excellence (DEW 2020). http://eval.how/dew2020/. (Accessed on 09/16/2020).

[9] Mahima Pushkarna, Andrew Zalvidar, and Dan Nanas. 2021. *Data Cards Playbook*.

[10] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Kumar Paritosh, and Lora Mois Aroyo. 2021. " Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. (2021).

[11] Marina Torre, Shinnosuke Nakayama, Tyrone J Tolbert, and Maurizio Porfiri. 2019. Producing knowledge by admitting ignorance: Enhancing data quality through an "I don't know" option in citizen science. *PloS one* 14, 2 (2019), e0211907.

[12] Meng-Han Wu and Alexander Quinn. 2017. Confusing the crowd: Task instruction quality on amazon mechanical turk. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 5.