“Everyone wants to do the model work, not the data work.” Data Cascades in High-Stakes AI

Summary and discussion by Ali Krzton

Those who work with data have learned the importance of provenance, documentation, standardization, context, and metadata in maintaining the quality of datasets. This was historically done to preserve their utility for human reuse and re-examination, but in recent years the emphasis on machine-readability of datasets has increased, in part to allow for their use in AI (artificial intelligence) applications. Just as those involved in creating and maintaining datasets benefit from an improved understanding of how they might be used with AI, the developers of AI systems should pay attention to issues that affect the data upon which their models rely. Several Google researchers present this perspective in “‘Everyone wants to do the model work, not the data work’: Data Cascades in High-Stakes AI”, a conference paper based on their qualitative study of AI practitioners (Sambasivan et al., 2021).

Sambasivan et al. (2021) interviewed 53 AI practitioners in East and West Africa, India, and the US who work on high-stakes AI, or AI applied in critical domains where failures have profound negative impacts on people. Through this work they identified a persistent problem, *data cascades*, originating from quality issues in the datasets used to build AI models rather than features of the models themselves. The report defines a data cascade as “compounding events causing negative, downstream effects from data issues, that result in technical debt over time” (Sambasivan et al., 2021: 5). That technical debt incurs human costs including harms to intended beneficiaries, abandonment of projects, alienation of project partners, and wasted time and effort (Sambasivan et al., 2021: 8).

While data cascades had multiple causes, the authors point to devaluation of data work with respect to model work as a central theme. This extended even to the domain expertise needed to understand and interpret the data in the first place, leaving the programmers to make classification and cutoff decisions they admitted they were not qualified to make (Sambasivan et al., 2021: 7). AI practitioners were aware that people involved with data collection and organization were not rewarded for their work, or if they were, they might be rewarded in ways that worked against data quality (Sambasivan et al., 2021: 9). Shortcomings in the education and training background of AI practitioners also contribute to data cascades. Most of them learned AI methodologies on extremely clean “toy” datasets or a selection of commonly used open datasets that were nothing like the real-world data on which they were required to build and train their models (Sambasivan et al., 2021: 11). Consequently, they were not prepared to deal with issues such as inaccurate, incomplete, non-representative, or poorly-documented data, leading to data cascades.

A canonical example of a high-stakes AI domain is healthcare. As AI tools are increasingly brought to bear on healthcare decisions, there is a growing risk that data cascades will lead to model problems discovered long after they have harmed substantial numbers of people. A separate study of the performance of a proprietary algorithm designed to provide early warning of sepsis by University of Michigan personnel found it flagged too many false positives while missing real cases of sepsis; the study’s lead author, Karandeep Singh, surmised that the model might be flawed because it was validated with billing codes rather than clinical data (Simonite, 2021). This is a data cascade, in this case arising from the use of data that was not a reliable indicator of the phenomenon of interest.

As AI researchers and practitioners alike discover the value of data work, data practitioners are presented with an opportunity to start new conversations and draw attention to the need for data expertise in AI-driven projects.

References

Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., and Aroyo, L.M. (2021). "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, May 8–13, 2021, Yokohama, Japan, 1-15. doi: 10.1145/3411764.3445518 Open access copy available at https://research.google/pubs/pub49953/

Simonite, T. (2021-06-21). An algorithm that predicts deadly infections is often flawed. *Wired*. https://www.wired.com/story/algorithm-predicts-deadly-infections-often-flawed/