Human-in-the-loop Bias Mitigation in Data Science

Anonymous Author(s)
Affiliation
Address
email

Abstract

With the successful adoption of machine learning (ML) in decision making, there have been growing concerns around the transparency and fairness of ML models leading to significant advances in the field of *eXplainable Artificial Intelligence* (XAI). Generating explanations using existing techniques in XAI and merely reporting model bias, however, are insufficient to locate and mitigate sources of bias. In line with the *data-centric AI* movement, we posit that to mitigate bias, we must solve the myriad data errors and biases inherent in the data, and propose a *human-machine* framework that strengthens human engagement with data to remedy data errors and data biases toward building fair and trustworthy AI systems.

1 Introduction

The data problem of AI. Algorithmic decision-making systems are increasingly being used to automate consequential decisions in a wide range of application domains such as healthcare, lending, hiring, and crime prevention and justice management. These systems are often touted to be amplifying existing societal biases and innocuous data errors that are reflected through the data the systems are trained upon [1, 2, 12]. To establish societal trust in machine learning (ML), the decisions generated by ML applications should be *robust* and *fair*, which mandates that the data used to built these applications be carefully evaluated and curated since multiple cases of ML applications violating human rights can be attributed to the low-quality data used for training the models [8].

Role of humans in shaping the data in AI. Human input is an important factor in machine learning pipelines. Researchers have long established that humans and their biases play an important role in data acquisition, data selection, curation, preparation and analyses [17]. These biases could be governed by social conditioning or be a result of unconscious cognitive propensities. It is, therefore, imperative to document the potential sources of human input but is often overlooked in addressing the fairness, transparency and explainability of machine learning models.

Human input in AI. In ML-based systems, human input is typically sought in the form of feedback from domain experts *after* the system generates outputs. While experts may interact with the ML model, they are rarely part of the design or development of the system itself. As an example, physicians in the domain of medicine routinely interact with systems but are not instrumental in their design and development. Building on the human-in-the-loop method [15], we consider human input in AI with respect to two dimensions: (1) role and impact of humans; (2) component of the data science pipeline. Specifically, the role of humans can be characterized by the type and amount of expertise the humans have. Domain experts/end users and designers have higher domain expertise but lower machine learning expertise. As a result, their input has lower impact on the AI system. On the other hand, as the *users* of the AI system, they *receive* higher impact from the AI system and vice versa for data scientists, data curators, and machine learning practitioners (Figure 1).

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

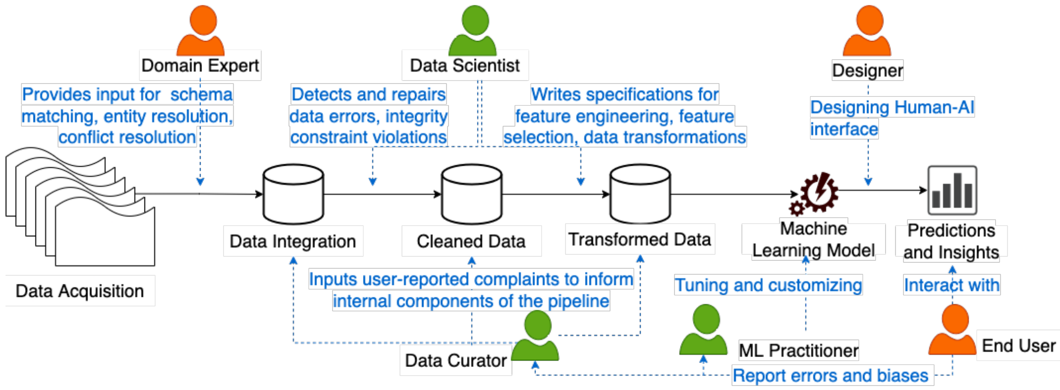

Figure 1: Framework showing human-machine interaction across the data science pipeline to improve effectiveness and fairness of the downstream ML task. The green and red colors represents different levels of domain/machine learning expertise and impact on/by the AI systems. Each profile icon represents a type of human roles, where different types of input are provided to the various components of the data science pipeline.

2 Strengthening human engagement with data

Ensuring high-quality data requires the ability to make informed data cleaning decisions (catering to different types of data errors and biases) at different parts of the ML application. This task requires coordination across the ML workflow so that data cleaning can account for downstream ML tasks, and the downstream parts can inform upstream cleaning decisions [13]. We propose to facilitate this coordination by strengthening human engagement with data in the ML pipelines.

Role of humans in “shaping" data. Integrating user feedback into the data science pipeline has been a much-studied area of research [10, 11]. Researchers, for example, have leveraged user feedback in consolidating heterogeneous data from multiple data sources for resolving entities [7, 9], matching data schema [14, 3], resolving conflicting information [4, 5, 6, 16], correcting data integrity errors [18] etc. Feedback is often sought on standalone pipeline components to improve their outcomes without much consideration to downstream data analyses.

The proposed human-machine integration (Figure 1) aims to characterize the influence of different human roles on the different components of the data science pipeline, identifying and resolving data errors in tandem with human input, thus facilitating trusted and fair ML. We seek to develop formal guidance on how to implement human-in-the-loop processes that facilitate robustness, and do not amplify or perpetuate the many human, systemic and computational biases that can degrade outcomes in the complex ML setting. We realize that it is, however, easier to ask users for feedback on the final output of the ML-based system (e.g., if the predictions made by the system are correct, fair) rather than on intermediate outputs (e.g., if a particular data curation step will lead to correct/fair outputs). In this context, we intend to highlight the power of human input along the data science pipeline by asking the following questions:

1. What is the right framework for soliciting human input for building fair and trustworthy AI systems?
2. How can we leverage human input in different components of the data science pipeline to resolve data-related issues and generate fair final decisions?
3. How can we design and prioritize questions to elicit meaningful human input with limited budget?
4. How can we incorporate noisy and uncertain human input and still guarantee fairness of the ML-based system?

We envision developing a framework that allows humans to inject knowledge at different stages of the data science pipeline, tracks the impact of those actions on the system decisions, and provides solutions to counter their potential harms on the society at large. Building such a framework requires designing new systems and developing data processing algorithms at the intersection of data management and human-computer interaction.

References

[1] Housing Department Slaps Facebook With Discrimination Charge. https://www.npr.org/2019/03/28/707614254/hud-slaps-facebook-with-housing-discrimination-charge, 2019.

[2] Self-driving Cars More Likely to Hit Blacks. https://www.technologyreview.com/2019/03/01/136808/self-driving-cars-are-coming-but-accidents-may-not-be-evenly-distributed/, 2019.

[3] AnHai Doan, Pedro Domingos, and Alon Y. Halevy. Reconciling Schemas of Disparate Data Sources: A Machine-Learning Approach. In *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*, SIGMOD '01, page 509–520, New York, NY, USA, 2001. Association for Computing Machinery.

[4] Wenfei Fan, Floris Geerts, Nan Tang, and Wenyuan Yu. Conflict Resolution with Data Currency and Consistency. *Journal of Data and Information Quality*, 5(1–2), Sep 2014.

[5] Wenfei Fan, Jianzhong Li, Shuai Ma, Nan Tang, and Wenyuan Yu. Towards Certain Fixes with Editing Rules and Master Data. *The VLDB Journal*, 21:213–238, 2011.

[6] Wenfei Fan, Shuai Ma, Nan Tang, and Wenyuan Yu. Interaction between Record Matching and Data Repairing. *Journal of Data and Information Quality*, 4(4), May 2014.

[7] Donatella Firmani, Barna Saha, and Divesh Srivastava. Online Entity Resolution Using an Oracle. *Proceedings of the VLDB Endowment*, 9(5):384–395, Jan 2016.

[8] European Union Agency for Fundamental Rights. Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights. 2019.

[9] Sainyam Galhotra, Donatella Firmani, Barna Saha, and Divesh Srivastava. Hierarchical Entity Resolution Using an Oracle. In *Proceedings of the 2022 International Conference on Management of Data*, SIGMOD '22, page 414–428, New York, NY, USA, 2022. Association for Computing Machinery.

[10] Shawn R. Jeffery, Michael J. Franklin, and Alon Y. Halevy. Pay-as-you-go user feedback for dataspace systems. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, page 847–860, New York, NY, USA, 2008. Association for Computing Machinery.

[11] Ashish Kapoor, Eric Horvitz, and Sumit Basu. Selective Supervision: Guiding Supervised Learning with Decision-Theoretic Active Learning. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, page 877–882, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[12] Matthew Kay, Cynthia Matuszek, and Sean A. Munson. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 3819–3828, New York, NY, USA, 2015. Association for Computing Machinery.

[13] Felix Neutatz, Binger Chen, Ziawasch Abedjan, and Eugene Wu. From cleaning before ml to cleaning for ml. *IEEE Data Eng. Bull.*, 44:24–41, 2021.

[14] Quoc Viet Hung Nguyen, Thanh Tam Nguyen, Zoltán Miklós, Karl Aberer, Avigdor Gal, and Matthias Weidlich. Pay-as-you-go Reconciliation in Schema Matching Networks. In *2014 IEEE 30th International Conference on Data Engineering*, pages 220–231, 2014.

[15] Besmira Nushi, Ece Kamar, Eric Horvitz, and Donald Kossmann. On human intellect and machine failures: Troubleshooting integrative machine learning systems. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[16] Romila Pradhan, Siarhei Bykau, and Sunil Prabhakar. Staging User Feedback toward Rapid Conflict Resolution in Data Fusion. In *Proceedings of the 2017 ACM International Conference on Management of Data*, SIGMOD '17, page 603–618, New York, NY, USA, 2017. Association for Computing Machinery.

[17] Reva Schwartz, Apostol Vassilev, Kristen Greene, Lori Perine, Andrew Burt, and Patrick Hall. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. In *Special Publication (NIST SP), National Institute of Standards and Technology, Gaithersburg, MD*, 2022-03-15 04:03:00 2022.

[18] Mohamed Yakout, Ahmed K. Elmagarmid, Jennifer Neville, Mourad Ouzzani, and Ihab F. Ilyas. Guided Data Repair. *Proceedings of the VLDB Endowment*, 4(5):279–289, Feb 2011.