Understanding and Mitigating the Negative Consequences of Training Dataset Explanations

Ariful Islam Anik
Andrea Bunt
University of Manitoba
Winnipeg, Manitoba, Canada
anikmai@myumanitoba.ca
andrea.bunt@umanitoba.ca

CHI 2023 Workshop on Human-centered Explainable AI (HCXAI), April 23–28, 2023, Hamburg, Germany.

Abstract
Owing to the importance of training datasets in the performance of AI systems, recent work in Explainable AI (XAI) has focused on communicating information about training datasets to stakeholders. While explaining this information can bring many potential benefits for the receivers, like other AI explanations, they can also bring negative consequences. In this position paper, we describe how we can use training dataset explanations to study the negative consequences of explanations and to explore potential mitigation strategies. We discuss potential challenges that researchers might face in these explorations.

Author Keywords
Training Dataset; Explanations; Machine Learning Systems; Presentation Style; Negative Impacts.

CSS Concepts
• **Human-centered computing~Human computer interaction (HCI)**; User studies;

Introduction
With the increasing use of black-box Machine Learning (ML) algorithms and Artificial Intelligence (AI) in automated decision-making systems, it has become

essential that stakeholders are educated and informed about these systems to ensure a good human-AI collaboration. To help stakeholders in this regard, transparency into these systems and how they work has been identified as of utmost importance [14,22,31]. To improve transparency, many explanation approaches have focused on providing information on how the systems make decisions and the factors that impact those decisions [2,6,10,12,16,35,36]. While these explanations can provide a range of benefits to the receivers (e.g., calibrated trust [9,15,25,27]), they can also bring unintentional negative consequences [13,21,24,29]. For example, sometimes the presence of explanations can hinder people's ability to identify system mistakes due to information overload [29]. Further, explanations can sometimes result in overtrust and overconfidence [21,24]. Prior work has defined these unintended negative outcomes as pitfalls of explanations [20].

One of the important components of ML systems is training datasets as the performance of these systems is highly dependent on the datasets and the training process [3,7,11]. To improve this aspect of system transparency, we have investigated how to communicate information on the training data and the training process to the stakeholders [1]. Our results thus far have been promising. Our studies have suggested that end-users perceive the explanations positively and that they have the potential to inform their trust and fairness judgments.

Despite the recent attention from the explainable AI (XAI) research community [1,6,18,19], there are open questions regarding explainability pitfalls. In the context of our studies with training dataset explanations, we saw indications of participants misinterpreting information on the training data and that this potentially impacted their perception of the system in an unintended way (e.g., participants interpreted the demographic distribution in a balanced way and perceived the system to be well-trained and positively when the demographic was imbalanced in many aspects). Training dataset explanations might also potentially lead to overconfidence in the system in situations where the training dataset was strong, but other performance metrics are weak.

In this position paper, we describe our approach to investigating potential negative consequences of training dataset explanations and generating ideas to minimize or mitigate them. We discuss how we can use training dataset explanations to potentially explore and gather insights into the impact of explanation presentation, the potential benefit of such explanations, and how they can result in negative consequences. Through these investigations, we aim to contribute to the XAI community's ongoing work on understanding what types of transparency can contribute to building effective human-AI partnerships.

Current State of Research on Training Dataset Explanations

To explore the idea of enhancing transparency by communicating training dataset information, we designed data-centric explanations [1] that provide information on training data (e.g., how the data was collected, the demographics of the data, and the recommended usage of the data). We used existing dataset documentation frameworks [23] to generate the information about a dataset and present it to the user via a Q&A-based approach. Through a user study,

we found that participants generally received this style of explanation positively and that the information impacted their trust and fairness judgments of the system.

Prior work in XAI has shown that information presentation in an explanation can impact receivers' perceived comprehension of the explanations [25,33], their understanding of the system [12], and their perceptions of system fairness [5]. Motivated by this, in an ongoing study, we are exploring information presentation in training dataset explanation and how it impacts receivers' understanding of it. In our exploration, we compare two different presentations for training dataset information - the existing Q&A-based approach [1,23] and a narrative-driven data story generated by following established properties of data storytelling [26,28,34]. In a between-subject study where participants critiqued an automated hiring system, we found that the presentation style of training dataset explanations influences what type of information participants focus on in their critique, their overall comprehension of the explanations, and impacts their perceptions of the system. However, we also noticed the possibility of participants misinterpreting some of the presented information which could impact their ability to develop calibrated trust. For example, some participants seemed to interpret the demographic distribution of a dataset as balanced even when this was far from the case. These types of misinterpretations motivate further study of how training dataset explanations might negatively impact the receivers.

Where Do We Go from Here?

Reflecting on the importance of training data in AI systems and the results of our initial investigation, the potential of training dataset explanations as an important measure of transparency in AI systems is evident. However, to utilize the benefits of the explanation, we need a better understanding of what impacts this type of explanation, how they are used by stakeholders, and how they might potentially result in negative consequences. Further, we need to be alert about the potential challenges in studying the explanation. In the following paragraphs, we provide more details about these directions for investigation.

Understanding the Benefits and Negative Impacts of Training Dataset Explanations

To gain insights into how design decisions made in training dataset explanations impact users' interaction with and understanding of the material, in our ongoing work, we are investigating the impact of presentation style. We have initial insights that receivers primarily focus on a subset of presented information and presentation style impacts their focus. Consequently, this potentially can negatively impact receivers' perception of the overall system as they might end up ignoring certain aspects of the information. Beyond presentation style, there are other design factors (e.g., comprehensiveness, modality) that can also impact the explanations. For example, due to the huge volume of information that can be provided about a training dataset, it is possible that training dataset explanation can lead to potential information overload and leave a negative impact on the receivers.

Based on these initial findings, we feel that it is important to gain a more systematic understanding of

how design factors of training dataset explanations impact the receivers and how we need to adapt the information for different contexts and stakeholders. For example, what is the right balance between the breadth and depth of the information being presented? Are there certain aspects of information that are important only in a context-specific way?

Identifying Strategies to Mitigate Potential Negative Consequences of Training Dataset Explanations
As outlined in the previous section, training dataset explanations can potentially lead to unintended negative consequences (pitfalls) for receivers. Further investigations are required to identify strategies that can mitigate the potential negative consequences. Since these pitfalls can appear based on how receivers perceive and interact with the explanation, one way of approaching this research is to involve AI and XAI practitioners. As supported by prior research, they are the ones who are most involved in the design of the explanations [4,17,30], and therefore, might be best suited to identify how pitfalls might appear in an explanation. They can also provide strategies on how to prevent or lessen the potential pitfalls. Further evaluations with the potential receivers can then assess the impact of these strategies on different stakeholders.

Challenges to Gathering Empirical Insights on How Stakeholders Use Training Dataset Explanations
While we have initial data suggesting that training dataset explanations are received positively by users [1], we do not know how different stakeholders might use training dataset explanations in practice. Moving towards more real-world deployments requires addressing several challenges. For example, owing to the volume of information presented, our initial evaluation suggested that training dataset explanations might be more important in high-stake domains than low-stakes domains [1]. However, studying the utility of the explanations in a high-stakes real-life context requires access to participants with specific skill sets, making recruiting challenging. Given these types of challenges, scenario-based studies [32] are often used to simulate high-stakes situations, however, these study designs lack ecological validity and can lead to evaluation pitfalls (e.g., proxy tasks, subjective measures) [8]. As such, there is a need for exploring the challenges with the evaluation of explanations, especially in high-stakes situations, to maximize the potential benefit of the explanations. For example, what are the challenges stakeholders face when using the explanation in practice? What is the most appropriate objective measure of evaluation?

Additionally, being a global explanation that provides information about training data, it is possible that training dataset explanations might create good initial impressions and help in the onboarding process. However, this could lead to an overreliance on the system when it comes to individual decisions. It is important to ensure that stakeholders can use such explanations over time. Therefore, there is also a need to gain an understanding of how training dataset explanations (or other types of explanations) are used by stakeholders over time.

Summary
In this position paper, we described our approach to investigating training dataset explanations as a type of system transparency. We also discussed the need for further work to understand and handle the potential

negative consequences of such explanations. We further touched on the potential challenges in studying this type of explanations and how we can work on maximizing the benefit of the explanations with the ultimate goal of having better human-AI collaboration.

References

1. Ariful Islam Anik and Andrea Bunt. 2021. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3411764.3445736

2. Liliana Ardissono, Anna Goy, Giovanna Petrone, Marino Segnan, and Pietro Torasso. 2003. Intrigue: Personalized recommendation of tourist attractions for desktop and hand held devices. *Applied Artificial Intelligence* 17, 8–9: 687–714. https://doi.org/10.1080/713827254

3. Solon Barocas and D Andrew. 2016. Selbst. 2016. *Big Data's Disparate Impact" California Law Review* 104: 671–732.

4. Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58: 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

5. Niels van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B. Skov. 2021. Effect of Information Presentation on Fairness Perceptions of Machine Learning Predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3411764.3445365

6. Reuben Binns, Max van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. "It's Reducing a Human Being to a Percentage": Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3173574.3173951

7. Buzz Blog Box. 2020. What is Training Data its types and why it is important? *Medium*. Retrieved from https://becominghuman.ai/what-is-training-data-its-types-and-why-it-is-important-f998424c3c9

8. Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, 454–464. https://doi.org/10.1145/3377325.3377498

9. Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-

making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1: 1–21. https://doi.org/10.1145/3449287

10. Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 258–262. https://doi.org/10.1145/3301275.3302289

11. Toon Calders and Indrė Žliobaitė. 2013. Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures. . 43–57. https://doi.org/10.1007/978-3-642-30487-3_3

12. Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–12. https://doi.org/10.1145/3290605.3300789

13. Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In *26th International Conference on Intelligent User Interfaces*, 307–317. https://doi.org/10.1145/3397481.3450644

14. A C M U S Public Policy Council. 2017. Statement on algorithmic transparency and accountability. *Commun. ACM*.

15. Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction* 18, 5: 455–496. https://doi.org/10.1007/s11257-008-9051-3

16. Anupam Datta, Shayak Sen, and Yair Zick. 2016. Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems. In *2016 IEEE Symposium on Security and Privacy (SP)*, 598–617. https://doi.org/10.1109/SP.2016.42

17. Advait Deshpande and Helen Sharp. 2022. Responsible AI Systems: Who are the Stakeholders? In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 227–236. https://doi.org/10.1145/3514094.3534187

18. Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 275–285. https://doi.org/10.1145/3301275.3302310

19. Upol Ehsan, Q. Vera Liao, Michael Muller, Mark O. Riedl, and Justin D. Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Conference on Human Factors in*

Computing Systems - Proceedings. https://doi.org/10.1145/3411764.3445188

20. Upol Ehsan and Mark O Riedl. 2021. Explainability pitfalls: Beyond dark patterns in explainable AI. *arXiv preprint arXiv:2109.12480*.

21. 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. https://doi.org/10.1145/3290607.3312787

22. Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing Transparency Design into Practice. In *23rd International Conference on Intelligent User Interfaces*, 211–223. https://doi.org/10.1145/3172944.3172961

23. Timnit Gebru, 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. https://doi.org/10.1145/3458723

24. Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376219

25. Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more?: the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1–10. https://doi.org/10.1145/2207676.2207678

26. Bongshin Lee, Nathalie Henry Riche, Petra Isenberg, and Sheelagh Carpendale. 2015. More Than Telling a Story: Transforming Data into Visually Shared Stories. *IEEE Computer Graphics and Applications* 35, 5: 84–90. https://doi.org/10.1109/MCG.2015.99

27. Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. *Why and why not* explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2119–2128. https://doi.org/10.1145/1518701.1519023

28. Adegboyega Ojo and Bahareh Heravi. 2018. Patterns in Award Winning Data Storytelling: Story Types, Enabling Tools and Competences. *Digital Journalism* 6, 6: 693–718. https://doi.org/10.1080/21670811.2017.1403291

29. 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. https://doi.org/10.1145/3411764.3445315

30. Alun Preece, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty. 2018. Stakeholders in explainable AI. *arXiv preprint arXiv:1810.00184*.

31. Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as Mechanisms for Supporting Algorithmic Transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3173574.3173677

32. Mary Beth Rosson and John M Carroll. 2009. Scenario-based design. In *Human-computer interaction*. CRC Press, 161–180.

33. Hong Shen, Haojian Jin, Ángel Alexander Cabrera, Adam Perer, Haiyi Zhu, and Jason I. Hong. 2020. Designing Alternative Representations of Confusion Matrices to Support Non-Expert Public Understanding of Algorithm Performance. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2: 1–22. https://doi.org/10.1145/3415224

34. Charles D Stolper, Bongshin Lee, Nathalie Henry Riche, and John Stasko. 2016. *Emerging and Recurring Data-Driven Storytelling Techniques: Analysis of a Curated Collection of Recent Stories*. Retrieved from https://www.microsoft.com/en-us/research/publication/emerging-and-recurring-data-driven-storytelling-techniques-analysis-of-a-curated-collection-of-recent-stories/

35. Paolo Tamagnini, Josua Krause, Aritra Dasgupta, and Enrico Bertini. 2017. Interpreting Black-Box Classifiers Using Instance-Level Visual Explanations. In *Proceedings of the 2nd Workshop on Human-In-the-Loop Data Analytics*, 1–6. https://doi.org/10.1145/3077257.3077260

36. Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces*, 318–328. https://doi.org/10.1145/3397481.3450650

1.