A Human-Centric Take on Model Monitoring

Murtuza N Shergadwala
Fiddler AI,
Palo Alto, CA, USA
murtuza@fiddler.ai

Himabindu Lakkaraju
Department of Computer Science,
Harvard University,
Cambridge, MA, USA
hlakkaraju@hbs.edu

Krishnaram Kenthapadi
Fiddler AI,
Palo Alto, CA, USA
krishnaram@fiddler.ai

Abstract

Predictive models are increasingly used for consequential decisions in high-stakes domains such as healthcare, finance, and policy. It becomes critical to ensure that these models make accurate predictions, are robust to shifts in the data, do not rely on spurious features, and do not unduly discriminate against minority groups. To this end, several approaches spanning various areas such as explainability, fairness, and robustness have been proposed in recent literature. Such approaches need to be human-centered as they cater to the understanding of the models to their users. However, there is a research gap in *understanding the human-centric needs and challenges of monitoring machine learning (ML) models once they are deployed*. To fill this gap, we conducted an interview study with 13 practitioners who have experience at the intersection of deploying ML models and engaging with customers spanning domains such as financial services, healthcare, hiring, online retail, computational advertising, and conversational assistants. We identified various human-centric challenges and requirements for model monitoring in real-world applications. Specifically, we found the need and the challenge for the model monitoring systems to clarify the impact of the monitoring observations on outcomes. Further, such insights must be actionable, customizable for domain-specific use cases, and cognitively considerate to avoid information overload.

1 Introduction

Machine Learning (ML) systems require maintenance not only by the virtue of possessing software code but also because of the nuances of ML as a domain itself [31]. ML-specific nuances include dependency of a model on (1) data and (2) outputs of another model. Data distributions can shift during production from when the model was designed. Further, the dependency of a model on another model's output can cascade issues from one model to another [31]. Such nuances imply that maintenance of ML systems requires *monitoring* its various aspects, such as data and models. More broadly, the emerging field of *MLOps* pertains to practices for deploying and maintaining ML models in production reliably and efficiently [22]. Further, monitoring ML systems is an essential part of a broader AI model governance [19] and responsible AI framework [2]. Such frameworks need to be human-centered not just in terms of usability by humans but also accounting for human behaviors [34, 39]. Refer Appendix A for more related work.

Key Contributions: Our study focuses on practical and real-world challenges by interviewing practitioners in industry. We identify various human-centered requirements and challenges in designing monitoring frameworks for ML systems. Moreover, we highlight the need to analyze such requirements and challenges from practitioners with caution as they are *perceived* challenges which can also stem from misconceptions about a systems' functionality.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

2 Study Design

We collected desiderata from thirteen (n=13) practitioners with expertise and/or interest in the MLOps space. We did so by conducting semi-structured and one-on-one interviews virtually and analyzed the results from each interviewee. All the interviewees had interacted with Fiddler AI's model monitoring platform and had an understanding about model monitoring. Based on an average of 2.7 years of professional experience of the interviewees within the MLOps space, we considered four to be experts, four to be beginners, and five to be intermediate. The interviews were semi-structured due to the exploratory and human-centered nature of the study which required prompting interviewees with follow-up questions relevant to their domain-specific experiences with model monitoring.

We asked the following questions. Q1) What kind of applications do you use ML models for? [To understand domain-specific use cases] Q2) Why do you need model monitoring in these applications? [To understand domain-specific desiderata] Q3) What aspects of model monitoring do you need? [To understand interpretations of model monitoring] Q4) What would an ideal model monitoring framework look like for you? [To understand human-centered desiderata for model monitoring]

We analyzed the interview responses using inductive content analysis [18]. The interview notes were analyzed twice. First, each transcript was individually annotated for model monitoring application use cases, requirements, and challenges described by each of the interviewees. Second, the responses to each question across all interviews were pooled and analyzed to inductively generate common themes and categories for ML monitoring application areas, design requirements, and challenges.

Responses to Q1 were identified either as a domain or a use case. For example, phrases and words such as "financial services", "insurance", "banking", and "adtech" were labeled as domains whereas phrases such as "fraud detection", "credit lending", and "speech recognition" were labeled as use cases. Responses to Q2, Q3, and Q4 were analyzed to identify human-centric desiderata for model monitoring. Sentences that contained phrases such as "need to", "should have", "be able to", and "it would be great if" were labeled as requirements. Phrases such as "difficulty", "challenges", "not possible", "hard", and "risky" were analyzed to identify challenges faced by the interviewees. Further, responses to Q3 were also analyzed to discover interviewees' interpretation of model monitoring and what it entails. To do so, authors' domain knowledge on model monitoring was leveraged to label the aspect of model monitoring discussed by an interviewee. For example, if interviewees discussed data drift as a part of the response to Q3, data drift was labeled as an aspect of model monitoring for that interviewee. The Q3 labels across all interviewees were pooled to characterize various aspects of model monitoring as discussed by them.

3 Application Areas, Design Considerations, and Challenges for Model Monitoring: Practitioners' Perspectives

We begin with the results of Q3 analysis, where we describe various aspects of model monitoring. Then, we discuss the results of Q1 analysis, namely, the application areas mentioned by the interviewees. Then, we discuss the interviewees' desiderata and challenges for model monitoring.

What is Model Monitoring? Practitioners' Perspectives: We analyzed the responses to Q3 contextually to identify the key aspects of model monitoring stated by the interviewees. All the interviewees discussed *model performance monitoring* and *data drift monitoring* as a part of model monitoring. Five (out of 13) interviewees emphasized *monitoring model fairness, bias, and model versions* as a part of model monitoring.

We also noticed interviewees categorizing or grading the relative importance of various aspects of model monitoring as a part of the response to Q3. Interviewees largely identify monitoring data integrity [4], that is, the accuracy, completeness, and consistency of data for inputs and outputs of a model as a *basic requirement*. In addition to this, identification of data drift, performance drift, and outlier detection were mentioned as important but *intermediate requirements* for model monitoring. Finally, monitoring model fairness and bias were expressed as regulation driven, "good to have", and *advanced requirements* for model monitoring.

Application Areas and Use Cases: Analyzing responses to Q1 enabled us to identify specific domains and use cases of interest to the interviewees as shown in Table 1. While there are numerous other domains and use cases for ML models, the results here are intended to contextualize the responses of the interviewees to the desiderata and the challenges discussed below.

Domain	Use Case
AdTech	Ads personalization and ads pricing.
Consumer Technology	Wake-word detection, automatic speech recognition, natural language understanding and interpretation, entity resolution, and text-to-speech generation.
Financial Services	Fraud detection, credit lending, and churn prediction.
Insurance	Risk prediction
Retail	Recommendation models and traffic monitoring.

Table 1: Domains and use cases discussed by the interviewees.

Human-Centric Requirements for Model Monitoring: The following themes for model monitoring requirements emerged from Q2, Q3, and Q4 responses.

Domain-Specific Debugging & Root Cause Analysis: Interviewees discussed the need for a model monitoring system to discover sub-populations of data where unexpected model behavior and outcomes occur to gain insight on model errors, when to retain a model, and domain-specific nuances. The system should also allow customizable levels of abstractions such as feature-level monitoring, prediction-level monitoring, and performance-level monitoring based on the use case.

Risk Management, Model Governance, and Privacy: Interviewees would like model monitoring systems to help them manage risk and ensure regulatory compliance. Interviewees emphasized the need for a monitoring system to enable centralization of model governance in an organization rather than have dependencies on an individual or a team that created the model. Such a system should also reduce human dependency and automate the process of error detection in ML pipelines. In certain settings, it would also be desirable for the monitoring system to ensure that privacy and confidentiality of various assets such as protected user information in training data and intellectual property associated with the models are protected.

Human-Centered Design: Interviewees discussed the need to preserve human autonomy and decision-making. In other words, the monitoring system should not trigger actions such as automated retraining. Instead, it should provide actionable suggestions to enable humans to make better decisions. The monitoring system should also provide relevant and meaningful alerts without cognitive overload. This implies system awareness of aspects such as the types of alerts, how often they are fired, and how they are presented.

Challenges for Model Monitoring in Practice: Interviewees highlighted several *practical challenges in designing and deploying an ML model monitoring system*. These challenges included design questions such as what should be monitored, how should the monitoring system interact with the model, and whether the monitoring system would also need the technical dependencies that a model requires, such as packages and modules, to ingest a model and execute successfully. These point to the requirement of a monitoring system having the ability to provide some guidelines or in-built options for what to monitor in an ML system. Moreover, monitoring systems are discussed as systems that have the ability to take an ML system as an input and provide monitoring insights as an output. Thus, the challenge for an ML monitoring system to possess a super-set of the technical capabilities of different ML systems is discussed. We note that this *perceived challenge* by the interviewees in practice is only a concern if the ML model is required to run specific predictions. We discuss this finding in Section 4.

Some interviewees had prior experience with model monitoring systems, and discussed the challenges of protecting privacy of users in training data and confidential information / intellectual property associated with their models. All interviewees emphasized the challenges of adapting a monitoring system to the domain-specific needs of the ML system and the lack of solutions that cater to their specific needs. For example, the data volume experienced by an ML system is highly sensitive to its application context, and could become a key design consideration for an ML monitoring system to be able to handle. Such decisions are currently made manually, and there is a lack of a framework that helps automate the design process for an ML monitoring system. Hence, interviewees highlighted the need for a fully managed monitoring service for their ML systems.

We also observed that the interviewees discussed *challenges that may occur when an ML monitoring system is deployed*. They discussed the lack of existing solutions in assessing whether an observed drift in data or model performance is a cause for concern. Such a lack of reliability can result in cognitive fatigue that may desensitize practitioners from gaining meaningful insights from the monitoring system. Interviewees also discussed latency challenges, that is, how quickly can a monitoring system detect issues and suggest remedial actions. Systems that are slow to identify issues may not be advantageous in certain contexts, such as autonomous driving, where high-stakes decisions may be made quickly by ML models. Further, due to privacy and IP issues, interviewees discussed the possibility of the monitoring tools and systems present "on-premise" or in-situ rather than a third party housing such a system. This implies that debugging or maintaining the monitoring system itself may be challenging, especially if the services are being sought from a third party.

Interviewees also discussed the *"so what" or value-based challenges* with monitoring systems. These discussions focused on the value of the insights gained from such systems, and pertained to aspects such as whether/how the monitoring insights could enable the stakeholders to take concrete actions to improve business outcomes and whether non-experts would be able to understand these insights. As noted earlier, data drift, outlier detection, data integrity violation, model performance, and bias/fairness are the key dimensions of model monitoring highlighted by the interviewees. Next, we discuss the specific desiderata themes within some of these dimensions.

For data drift, we found that all the interviewees consider input and output data distribution monitoring as a necessity. Further, such drift monitoring is treated as an indicator for model retraining. Further, data integrity violation and outlier detection are aspects that are discussed as a part of the data drift monitoring functionality as well. For bias/fairness monitoring, customer requirements are currently driven through policies and regulations. Words such as "policy","regulatory constraints","fines","regulatory push" were used to describe the need for fairness monitoring. Practitioners mentioned that compliance and risk management teams are concerned about bias/fairness also due to its impact on the trust and the reputation of a company. In this manner, an organization could proactively detect and mitigate any biases observed in its deployed models instead of having to react when such issues are discovered by external entities.

4 Conclusion and Discussion

Motivated by the need for understanding the human-centered requirements and challenges in designing ML systems monitoring frameworks, we performed an interview study with ML practitioners with experience spanning several application domains. We presented findings and insights on real-world use cases, desiderata, and challenges for ML model monitoring in practice based on these interviews. Interviewees discussed both feature-specific and process-specific aspects of model monitoring. Feature-specific aspects include monitoring data drift, model performance, and bias/fairness and ensuring that the alerts are relevant without cognitive overload. Process-specific aspects include the temporal considerations before, during, and after the deployment of the model monitoring system and the ability of a monitoring system to cater to different needs across the lifecycle of and ML system.

We highlight a *perceived challenge* on the technical requirements of a model monitoring system. Interviewees described the challenge for an ML monitoring system to have a super-set of the technical capabilities of different ML systems it monitors. However, we note that to monitor inputs, outputs, and other characteristics of a deployed model, a monitoring system may not need to execute the model. Instead, one might be able to take the production logs associated with the model as input. Thus, the model monitoring system may not need the technical infrastructure to run the model itself, and can instead focus on the *tools and techniques for prediction of distribution shifts*. This understanding influences the design decisions for monitoring systems and we thus highlight that model monitoring systems may be designed in a model-agnostic manner. This also points to the caution required in analyzing human-centric desiderata where perceived challenges by practitioners, who may not necessarily be experts in ML, may stem from misconceptions about the system functionalities.

We acknowledge that our analysis is limited based on the inputs of only thirteen practitioners. Future work includes conducting surveys and analyzing perspectives on model monitoring from a broader population of the ML practitioner community. Moreover, we encourage MLOps practices to formalize design frameworks for ML monitoring systems that are cautiously informed by human-centered desiderata.

Acknowledgments and Disclosure of Funding

Acknowledgements

We thank the thirteen practitioners for volunteering and providing their insights about model monitoring. We would also like to thank Joshua Rubin and Lea Genuit for providing feedback on the paper and our analysis.

References

[1] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen, et al. Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2019.

[2] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115, 2020.

[3] Azure. MLOps: Model management, deployment, lineage, and monitoring with Azure Machine Learning, 2022.

[4] J. E. Boritz. Is practitioners' views on core concepts of information integrity. *International Journal of Accounting Information Systems*, 6(4):260–279, 2005.

[5] E. Breck, N. Polyzotis, S. Roy, S. Whang, and M. Zinkevich. Data validation for machine learning. In *MLSys*, 2019.

[6] J. M. Carroll. Human-computer interaction: Psychology as a science of design. *Annual review of psychology*, 48(1):61–83, 1997.

[7] M. Cooley. Human-centered design. *Information design*, pages 59–81, 2000.

[8] G. Cormode, Z. Karnin, E. Liberty, J. Thaler, and P. Veselỳ. Relative error streaming quantiles. In *PODS*, 2021.

[9] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[10] Fiddler. Explainable monitoring, 2022.

[11] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys (CSUR)*, 46(4):1–37, 2014.

[12] S. Garg, Y. Wu, S. Balakrishnan, and Z. Lipton. A unified view of label shift estimation. In *NeurIPS*, 2020.

[13] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[14] M. Hardt, X. Chen, X. Cheng, M. Donini, J. Gelman, S. Gollaprolu, J. He, P. Larroy, X. Liu, N. McCarthy, A. Rathi, S. Rees, A. Siva, E. Tsai, K. Vasist, P. Yilmaz, M. B. Zafar, S. Das, K. Haas, T. Hill, and K. Kenthapadi. Amazon SageMaker Clarify: Machine learning bias detection and explainability in the cloud. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2021.

[15] J. Hermann and M. D. Balso. Meet Michelangelo: Uber's machine learning platform, 2017.

[16] IBM. Validating and monitoring AI models with Watson OpenScale, 2022.

[17] Z. Karnin, K. Lang, and E. Liberty. Optimal quantile approximation in streams. In *FOCS*, 2016.

[18] K. Krippendorff. *Content Analysis: An Introduction to Its Methodology (Commtext Series)*. Sage Publications, Thousand Oaks, California, 1980.

[19] E. Kurshan, H. Shen, and J. Chen. Towards self-regulating AI: Challenges and opportunities of AI model governance in financial services. In *Proceedings of the First ACM International Conference on AI in Finance*, 2020.

[20] Z. Lipton, Y.-X. Wang, and A. Smola. Detecting and correcting for label shift with black box predictors. In *ICML*, 2018.

[21] M. Maguire and N. Bevan. User requirements analysis. In *IFIP World Computer Congress, TC 13*, pages 133–148. Springer, 2002.

[22] S. Mäkinen, H. Skogström, E. Laaksonen, and T. Mikkonen. Who needs mlops: What data scientists seek to accomplish and how can mlops help? In *2021 IEEE/ACM 1st Workshop on AI Engineering-Software Engineering for AI (WAIN)*, pages 109–112. IEEE, 2021.

[23] C. F. Martin. *User-centered requirements analysis*. Prentice-Hall, Inc., 1988.

[24] K. P. Murphy. *Machine learning: A probabilistic perspective*. MIT press, 2012.

[25] D. Nigenda, Z. Karnin, M. B. Zafar, R. Ramesha, A. Tan, M. Donini, and K. Kenthapadi. Amazon SageMaker Model Monitor: A system for real-time insights into deployed machine learning models. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022.

[26] D. A. Norman. How might people interact with agents. *Communications of the ACM*, 37(7):68–71, 1994.

[27] S. Reddi, B. Poczos, and A. Smola. Doubly robust covariate shift correction. In *AAAI*, 2015.

[28] M. O. Riedl. Human-centered artificial intelligence and machine learning. *Human Behavior and Emerging Technologies*, 1(1):33–36, 2019.

[29] D. Sacha, M. Sedlmair, L. Zhang, J. A. Lee, D. Weiskopf, S. North, and D. Keim. Human-centered machine learning through interactive visualization. In *ESANN*, 2016.

[30] S. Schelter, D. Lange, P. Schmidt, M. Celikel, F. Biessmann, and A. Grafberger. Automating large-scale data quality verification. In *VLDB*, 2018.

[31] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison. Hidden technical debt in machine learning systems. In *NeurIPS*, 2015.

[32] M. N. Shergadwala and M. Seif El-Nasr. Human-centric design requirements and challenges for enabling human-ai interaction in engineering design: An interview study. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 85420, page V006T06A054. American Society of Mechanical Engineers, 2021.

[33] B. Shneiderman. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6):495–504, 2020.

[34] B. Shneiderman. Responsible AI: Bridging from ethics to practice. *Communications of the ACM*, 64(8):32–35, 2021.

[35] A. Taly, K. Sato, and C. Gruia. Monitoring feature attributions: How Google saved one of the largest ML services in trouble, 2021. Google Cloud Blog.

[36] A. Tsymbal. The problem of concept drift: definitions and related work. *Computer Science Department, Trinity College Dublin*, 106(2):58, 2004.

[37] L. Wasserman. *All of statistics: A concise course in statistical inference*, volume 26. Springer, 2004.

[38] G. I. Webb, R. Hyde, H. Cao, H. L. Nguyen, and F. Petitjean. Characterizing concept drift. *Data Mining and Knowledge Discovery*, 30(4):964–994, 2016.

[39] J. M. Wing. Trustworthy AI. *Communications of the ACM*, 64(10):64–71, 2021.

[40] Y. Wu, E. Winston, D. Kaushik, and Z. Lipton. Domain adaptation with asymmetrically-relaxed distribution alignment. In *ICML*, 2019.

[41] W. Xu. Toward human-centered AI: A perspective from human-computer interaction. *Interactions*, 26(4):42–46, 2019.

[42] I. Žliobaitė, M. Pechenizkiy, and J. Gama. An overview of concept drift applications. *Big data analysis: new algorithms for a new society*, pages 91–114, 2016.

A Appendix: Related Work

Techniques & tools: There is a rich literature on techniques for drift detection and model monitoring (e.g., see [17, 8, 5, 12, 20, 27, 40, 11, 38, 36, 42, 37, 24, 13, 9] and the references therein). There are also several open source and commercial tools for monitoring deployed ML models, e.g., Amazon SageMaker Model Monitor [25] & Clarify [14], Deequ [30], Fiddler's Explainable Monitoring [10], Google Vertex AI Model Monitoring [35], IBM Watson OpenScale [16], Microsoft Azure MLOps [3], and Uber's Michelangelo platform [15]. In contrast to these techniques and tools, our focus is on understanding of needs and requirements of the human stakeholders who ultimately leverage them.

Human-centered design: The notion of involving the humans who will ultimately use and interact with ML systems is not new [29]. Domains such as Human-Centered Design (HCD) and Human-AI Interaction (HAI) focus on design principles that enable humans to interact with intelligent and interactive tools and agents [26]. The area of HCD [7] has already established the need for the design of computational tools and techniques in a manner that satisfies the needs and requirements of the humans using them [21, 23]. However, the rise of new AI technologies for MLOps such as monitoring systems requires revisiting HCD principles and contextualizing them to the design of AI [28]. Existing literature offers guidelines for HAI interaction design from the perspective of human factors [1], engineering design [32], psychology [6], human computer interaction [41], and trustworthy AI [33]. However, in the context of ML model monitoring, there is a lack of understanding of the needs of the human stakeholders towards leveraging such HCD principles.