Understanding Practitioners' Challenges in Evaluating UX Outcomes of AI/ML applications

MEENA DEVII MURALIKUMAR, Human-centered Design and Engineering, University of Washington, , USA
DAVID W. MCDONALD, Human-centered Design and Engineering, University of Washington, , USA

In this paper, we describe our academic work that focuses on understanding the challenges and needs of UX practitioners who work with Artificial Intelligence (AI) and Machine Learning (ML). Specifically, for the purposes of the workshop, we present challenges discussed by UX practitioners that relate to 1) evaluating AI/ML applications from a user-centered perspective and 2) tackling the large-scale nature of AI/ML applications, especially post deployment. We also summarize our prior academic work on evaluating publicly available AI and ML models from a human-centered perspective and discuss how we might apply our tried and tested research methods towards developing solutions for UX practitioners.

Additional Key Words and Phrases: UX practitioners, evaluation, AI/ML, UX metrics

ACM Reference Format:
Meena Devii Muralikumar and David W. McDonald. 2023. Understanding Practitioners' Challenges in Evaluating UX Outcomes of AI/ML applications. 1, 1 (August 2023), 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In this workshop paper, we consider how UX practitioners understand and evaluate user-centered outcomes for AI/ML products and services. We discuss data from our recent interview study with UX practitioners on how they work with AI/ML and the challenges they face. Participants in our study discussed different kinds of challenges related to collaboration with data scientists, the organizational context, and the prototyping and evaluation processes. We wanted to highlight challenges of conducting user-centered evaluation of AI/ML products and how they relate to the goals of the workshop.

Though the domain area is specific to AI/ML applications, many of the issues described by our participants echoed broader challenges and necessities of evaluating UX metrics, such as getting stakeholder buy-in, communication and collaboration with cross-functional team members, and dealing with prioritization of business metrics over UX metrics [1]. Rather than shift the focus to the peculiarities of AI/ML, we hope to understand how different challenges in evaluating UX outcomes at scale, generally, also affects evaluation practices of AI/ML systems for UX practitioners.

The rest of the paper is structured as follows. First, we discuss data from UX practitioners working on different kinds of AI/ML products and services. Evaluating user-centered outcomes and tackling the scale at which most AI/ML applications function were the two key challenges described by participants. Next, we describe our prior work in conducting human-centered evaluations of AI/ML models and motivations to translate our methods from research to

Authors' addresses: Meena Devii Muralikumar, Human-centered Design and Engineering, University of Washington,, USA, mmeena@uw.edu; David W. McDonald, Human-centered Design and Engineering, University of Washington,, USA, dwmc@uw.edu.

practice. We conclude with ideas for future work aimed at building tools and supporting practitioners in conducting UX evaluations of AI/ML products.

2 EVALUATING AI/ML AND TACKLING SCALE

In this section, we recount our participants' experience of evaluating AI and ML based products and services. One of the main challenges that UX practitioners faced with respect to working on AI/ML was evaluating it from a user-centered perspective. UX practitioners pointed out the gulf between evaluating model-centric measures (accuracy, precision, confidence scores) and user-centric outcomes and the need to reconcile them.

> *This is actually more of a personal opinion on how ML data scientists work. They are looking at the model performance. They are not looking at the qualitative side of the impact. And so there is a big gulf there. Just because your model is performing correctly doesn't mean that users are, you know, actually enjoying the thing that you are showing them, right? That you have nailed it? You can describe it as looking at outcomes versus outputs.* - P3

Other UX practitioners also engaged in similar efforts to influence a human-centered approach on how AI/ML models are evaluated and to ensure that they are evaluated according to the context of use. For example, P5 posed the question,

> *Which parts are really cool for science and academia and which parts are actually going to be useful to the users?*

However, they were not always successful in these efforts. As P9 points out, the models developed by ML practitioners or researchers did not always satisfy the needs of their product users.

> *If there's a new research project or proposal coming up, we're hoping that it's motivated by some needs or pain points that we have found previously. Instead of doing it the other way around - like building the model and then finding out that it's not as useful as we hoped.* - P9

An important challenge and need to be addressed for these UX practitioners is in regard to the methods and tools for evaluating the AI/ML component from a user-centered standpoint. However, the scale at which most AI/ML systems operate posed challenges for them. Consider this anecdote from P3 (paraphrased),

> *I was one of two people who read the chatbot transcripts. It was a lot of contextual analysis, thematic coding… At one point I discovered that the algorithm is interpreting some information - some answers wrong - leading to unfair outcomes. This was potentially a serious situation as it had real-life consequences for those interacting with the chatbot. I only caught it because I was looking at the outcomes. And there was one of me looking at 100 conversations a day. And we were doing tens of thousands of conversations in a day. That is what I mean by scale. These are all built for scale.* - P3

P3 liked to leverage qualitative methods to stay close to how the technology performed. In this case, P3 was somewhat successful in ascertaining the chatbot's impact and its problematic points. But they could not manually handle the huge number of transcripts generated by the chatbot everyday. Addressing the challenges of evaluating UX for these kinds of large-scale AI/ML applications is a necessary first step to better tackle important issues of fairness and bias as well. P14 noted how as UX Researchers they *'get pulled into business goals'*, resulting in product or other metrics such as task completion subsuming evaluation of the AI/ML feature.

We must also consider how the model continues to learn and evolve post deployment. Indeed, when Yang et al. [5] attributed one source of AI's design complexity to its *capability uncertainty* [1], they considered how learning new data after deployment can cause positive or negative consequences depending on the user, context, and the new data. This aspect further stresses the need to evaluate and monitor AI/ML models from a UX perspective. In earlier stages of development, UX practitioners might neither be able to simulate the full range of model capabilities, nor can they predict how it will evolve with new data post deployment.

P3 also discussed the challenge of not having the right tools or techniques to see what end-users are being shown and how they are accepting the model outputs.

> *One of the challenges I see regularly in my space is, we don't have the tools to test the algorithm, the output of the algorithms. Again - I can see it has a 95% confidence rate. Right? It is working and that. But I don't know what people are being shown. So, and it depends - I am generalizing it a lot but we don't always have the tools to see how our model is performing in production environments.* - P3

P9 discussed a different aspect of scale in evaluating AI/ML applications such as recommender systems where a longitudinal study would be effective but not necessarily feasible to implement.

> *I think recommendations usually get better if a user interacts with it for a longer period of time - giving feedback, tuning it over time. So it's hard to ask them to share how useful it is just based on one interaction. So maybe a longitudinal study or over time study will be better. I don't know how feasible that is. It may require a lot of work before the study - like we have to create a separate prototype and record all the interactions. Things like that.* - P9

This account raises further questions of how UX practitioners working on AI/ML systems account for changes in model performance over time. Longitudinal user research studies seem warranted but it is unclear whether it is common practice, what specific metrics are used for measuring algorithm impact over time, and if the research infrastructure exists to conduct such studies. When probed about needs to address such challenges, P3 surfaced the idea of a 'testing tool' that can potentially help investigate model outputs and its rationale.

> *They are not updating the models all the time. You build the model, you deploy it, at what point do you commit to working on it again? I think that's where the testing tools would be really valuable. I would love a tool that gives me the information so I can question it. It's almost like stopping the model and saying 'Okay! How did you come up with this?' - and maybe it gets into explainability and stuff like that… And then see what the customer clicked on, what choices did they make, how are they using it… Because then you have an actual visual and you can pinpoint it [to the data scientist] versus looking at the performance of the model itself through a quantitative lens.* - P3

Making the model outputs and its rationale visible, to whatever extent possible, for investigation by the UX practitioner is an important first step. As P3 explains, these kinds of tools and information are required to check if the 'accurate model' also meets user needs and get stakeholder buy-in for making model updates in case there are discrepancies.

[1] Capability uncertainty of AI refers to how a model's capabilities cannot be fixed or predetermined throughout the different stages in its lifecycle - given its probabilistic nature.

2.1 Key Takeaways

(1) For UX practitioners, the introduction of AI/ML models in different products and services has raised the issue of how to formulate and measure user-centered outcomes that complement model-centric metrics, and highlight its importance to different stakeholders.

(2) UX practitioners are facing challenges in tackling the scale and output complexities [5] of the AI/ML model, especially post deployment when the model continues to learn with user data. Prior work has also called attention to the need of adapting design and UX methods to better monitor and evaluate AI/ML models post launch [6].

Though we might need more data to support it, we believe that the participants we interviewed are caught in a catch-22 type of situation where they need data to convince stakeholders about the need to develop methods, metrics, and techniques to evaluate the UX of an AI/ML system. But they need the tools and/or technical support to collect that data related to model interaction and use in the first place.

3 CONDUCTING HUMAN-CENTERED EVALUATIONS FOR AI/ML MODELS

In this section, we summarize academic work that the authors have engaged in previously to evaluate publicly available AI and ML models from a human-centered perspective. In [2], an ML tool that labels politeness (polite, neutral, impolite) was revalidated by collecting user ratings of politeness on data very similar to what the original classifier was trained on. Though the politeness classifier had a high, overall accuracy rate, Hoffman et al. [2] found that the tool was more effective in classifying categories of polite text than at classifying impolite or neutral categories of text.

Similarly, in [4] the authors evaluated Perspective - a publicly available ML-based API for toxicity detection. They conducted an human-centered evaluation by investigating if Perspective's toxicity scores aligned with user ratings of toxicity across data from three different platforms. They also checked if the classifier had any latent attributes by asking users to rate formality, respectfulness, and presence of stereotypes and understanding how these attributes varied with Perspective's toxicity scores. They found that a high toxicity score from Perspective aligned with user ratings for toxicity and disrespectfulness across all three platforms. However, it also aligned with user ratings for informality and high presence of stereotypes, indicating the possibility of latent attributes.

In both these studies, the same statistical test is used to draw inferences about how human judgements of textual attributes correspond to machine judgements of the same attributes - the logistic regression test [3]. In [2], nominal categories are used for politeness and hence the multinomial logistic regression is used whereas in [4], the ordered nature of toxicity labels are considered and the ordered logistic regression is used. We consider our evaluations to be human-centered because the fundamental question that the statistical test is modeled after is - *How well do humans agree with the predictions of an AI/ML tool?* As pointed out in [4], the direction of our question was aimed at understanding how the AI/ML tool might satisfy end-user expectations post training or deployment, rather than during the creation of the AI/ML tool itself (where we might be more interested in how the model agrees and follows user data or training data). Also, our methods allowed us to draw inferences about how the AI/ML model would perform in different sociotechnical contexts.

We find ourselves applying our experience from doing this kind of research towards practitioners' challenges of evaluating AI/ML to ask the following questions:

(1) How might we leverage our methods to help devise user-centered metrics for AI/ML evaluations? Could we potentially use the suite of logistic regression tests available to tackle the scale complexities of AI/ML applications?

(2) What are the opportunities and limitations of using behavioral log data (apart from using explicitly collected data or crowdsourced data) to run these statistical tests and draw broad inferences about user feedback on the AI/ML model?

(3) What organizational, technical, and methodological challenges hinder application of such statistical tests towards evaluating UX outcomes for practitioners?

One direction of future work that we have been considering is to develop an open-source toolkit that potentially elevates challenges related to the knowledge and implementation of these statistical tests. Another valuable direction of future work is to investigate how the BLUE framework [1] can be used to support practitioners in developing methods and metrics for user-centered, large-scale evaluations of AI/ML products.

4 CONCLUSION

In this workshop paper, we have discussed challenges UX practitioners face in evaluating large-scale AI/ML applications from a user-centered perspective. Also, we offer implications of our prior work in evaluating AI/ML models from a human-centered standpoint towards addressing practitioners' challenges. We hope both these topics will generate meaningful and relevant discussions for the workshop. We also hope to understand and learn from perspectives of industry practitioners and researchers to inform our ongoing work on this topic.

REFERENCES

[1] Serena Hillman, Samira Jain, Vichita Jienjitlert, and Paula Bach. 2022. The BLUE Framework: Designing User-Centered In-Product Feedback for Large Scale Applications. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 21, 8 pages. https://doi.org/10.1145/3491101.3503558

[2] Erin R. Hoffman, David W. McDonald, and Mark Zachry. 2017. Evaluating a Computational Approach to Labeling Politeness: Challenges for the Application of Machine Classification to Social Computing Data. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 52 (dec 2017), 14 pages. https://doi.org/10.1145/3134687

[3] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. 2013. *Applied logistic regression*. Vol. 398. John Wiley & Sons.

[4] Meena Devii Muralikumar, Yun Shan Yang, and David W. McDonald. 2023. A Human-Centered Evaluation of a Toxicity Detection API: Testing Transferability and Unpacking Latent Attributes. *Trans. Soc. Comput.* 6, 1–2, Article 4 (jun 2023), 38 pages. https://doi.org/10.1145/3582568

[5] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-Examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376301

[6] Sabah Zdanowska and Alex S Taylor. 2022. A Study of UX Practitioners Roles in Designing Real-World, Enterprise ML Systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 531, 15 pages. https://doi.org/10.1145/3491102.3517607