Accuracy is Not All You Need

David Piorkowski
IBM Research
Yorktown Heights, NY
djp@ibm.com

Rachel Ostrand
IBM Research
Yorktown Heights, NY
rachel.ostrand@ibm.com

Yara Rizk
IBM Research
Cambridge, MA
yara.rizk@ibm.com

Vatche Isahagian
IBM Research
Cambridge, MA
vatchei@ibm.com

Vinod Muthusamy
IBM Research
Austin, TX
vmuthus@us.ibm.com

Justin D. Weisz
IBM Research
Yorktown Heights, NY
jweisz@us.ibm.com

Abstract

Improving the performance of human-AI (artificial intelligence) collaborations tends to be narrowly scoped, with better prediction performance often considered the only metric of improvement. As a result, work on improving the collaboration usually focuses on improving the AI's accuracy. Here, we argue that such a focus is myopic, and instead, practitioners should take a more holistic view of measuring the performance of AI models, and human-AI collaboration more specifically. In particular, we argue that although some use cases merit optimizing for classification accuracy, for others, accuracy is less important and improvement on human-centered metrics should be valued instead.

1 Introduction

Human-AI collaboration, in which a human and an Artificial Intelligence (AI) model work together to complete a task, has potential for big improvements in performance over either entity's work alone. For example, AI components can aid a human in decision making tasks (e.g., [10, 11]), automating tedious work (e.g. [15]), or generating artifacts such as stories or source code (e.g., [6, 17, 19]).

AI model performance, and, by extension, human-AI collaborations, are often only evaluated using measures of the AI's classification or prediction performance, which can lead to practitioners chasing after (sometimes miniscule) performance gains to nudge out the current state-of-the-art (e.g., [1]). However, for many use cases, human-AI collaborations would benefit from being evaluated not solely on classification performance, but on user-centric measures as well.

There is often little appetite for improving systems along other dimensions, even though human-AI collaborations are highly multidimensional. We argue that it is important for "performance improvements" to also include measures at the interface between the human and the AI – both quantitative and qualitative – that encompass aspects of the user experience and the collaborative process. In doing so, other avenues for improving AI models open up, allowing system builders to focus on different measures based on the use case of their model, and allowing for a measurable improvement even when gains to the AI's prediction performance are difficult or prohibitive.

2 Three Dimensions of Performance Measures

The goal of a human-AI collaboration is to enable a user and an AI system to perform a task more effectively together than either entity could do alone. However, there are multiple ways to define

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

effectiveness, and different situations or types of AI systems call for optimizing on different measures. Here, we lay out three dimensions of measures on which an AI system could be evaluated — and thus, optimized for — to argue that it is important for system builders to take into account which measure(s) is the most impactful for their particular use case, and selectively optimize for that one.

We define and illustrate the three dimensions using the example of an end user who employs a generative AI model to help translate a piece of code from Java to Python (motivated by [17]).

AI model measures focus on the prediction performance and/or quality of the underlying AI model(s). Specific metrics include *error rates* (e.g., accuracy, F1-score, false positives, mean squared error, ROC curves), *computational efficiency* (e.g., training/prediction time), *number of training samples*, and *model size* (e.g., number of hyper-parameters in a neural network). For the code translation example, AI measures also include the number of errors that need to be corrected in the source code before that code passes a set of unit tests, the rate of errors, and the time taken by the model to produce the translation [2].

Interaction measures focus on the efficiency of the interaction between a human and the AI system. Metrics include the number of steps to perform a task, task completion time for the user, and the number of errors or digressions made during the joint task. These metrics are mostly quantitative and often can be captured automatically via logging within the software. For the code translation example, Interaction measures include the number of additional edits the user made to the AI's translation, the time to finish the task, and the number of compiler and linting errors the programmer corrected.

User Experience (UX) measures focus on the human's experience using the AI system. UX evaluations can contextualize Interaction measures by capturing qualitative perceptions and feedback, as well as quantitative measures, often to judge the effect of a particular feature or design. Quantitative UX metrics include measures of cognitive load [8], usability [12, 7, 4, 14], creativity support [5], trust [9], affect [13, 16], perceptions [3], and acceptance [7]. For the code translation example, UX measures include the user's effort in producing a correct translation, the extent to which they felt productive on the task, and the extent to which they were satisfied with their work.

By expanding the measurement of AI performance, especially in the context of human-AI collaboration, to include metrics from these latter two dimensions, new opportunities for enabling users to be more effective emerge. By assessing Interaction measures, improvements could subsequently be made to reduce the number of steps needed to complete a task or provide additional support to help understand the system, thereby reducing human errors. By assessing UX measures, user feedback could then be used to help identify pain points in the system that are not necessarily caused by the AI components. Even systems with the highest prediction accuracy have limited utility if the user interface is impossible to interact with, or imposes an unreasonable cognitive load on the practitioner.

It is additionally useful to consider the relative importance between these different measures for a given system, and the trade-off between them can vary from one system to another. A system that classifies social media posts would likely benefit from high performance on usability and cognitive load metrics, whereas top-notch prediction accuracy is less critical. In contrast, for an AI system whose goal is to diagnose cancer, optimizing accuracy as much as possible (without, of course, falling prey to overfitting on artifacts in the training set [18]), is likely the most important performance metric – as long as Interaction and UX measures are sufficiently good for users to be willing and able to interact with the system. Thus, in addition to improving the AI, improving the efficiency of the interaction and providing a better user experience can net improvements in the overarching goal: more effective human-AI collaboration that ultimately leads to more adoption and real-world impact.

3 Conclusion

The goal of human-AI collaboration is to enable AI solutions for real-world applications. Optimizing an AI model solely for prediction accuracy leads to missed opportunities in improving its holistic performance, and may not even be the most important criterion to focus on for a particular system. Even the most accurate system is worthless if people are unwilling or unable to use it because it takes too long to run, there are too many steps required of the user, it is perceived as untrustworthy, or the interface is too complicated to understand. Similarly, it is a waste of resources to work on improving the AI's prediction accuracy by an amount that is too small for the user to notice any difference in performance, or because interaction issues prevent the improvements from being detected by the user.

We call for the AI and HCI communities to adopt a more holistic approach to measuring performance by considering different dimensions of the AI system, and not just focusing on its prediction performance. We suggest measuring a system along three dimensions – AI model, Interaction, and UX – early and often, and to keep them in mind during the design process. Practitioners will only optimize what they can measure, and will only measure what they think to be of value. By sharing these dimensions with the community, we hope to spread acceptance of careful consideration of what dimensions are most relevant for each particular system and spur research to guide others in leveraging these dimensions effectively.

References

[1] Speech Recognition on LibriSpeech test-clean. `https://paperswithcode.com/sota/speech-recognition-on-librispeech-test-clean`. Accessed: 2022-09-26.

[2] Mayank Agarwal et al. Quality estimation & interpretability for code translation. *arXiv preprint arXiv:2012.07581*, 2020.

[3] Joey Benedek and Trish Miner. Measuring desirability: New methods for evaluating desirability in a usability lab setting. *Proc. Usability Professionals Association*, 2003(8-12):57, 2002.

[4] John Brooke et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7, 1996.

[5] Erin Cherry and Celine Latulipe. Quantifying the creativity support of digital tools through the creativity support index. *ACM Trans. Computer-Human Interaction*, 21(4):1–25, 2014.

[6] Elizabeth Clark et al. Creative writing with a machine in the loop: Case studies on slogans and stories. In *23rd Int. Conf. Intelligent User Interfaces*, pages 329–340, 2018.

[7] Fred D. Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13:319–340, 1989.

[8] Sandra G Hart. NASA-task load index (NASA-TLX); 20 years later. In *Proc. human factors and ergonomics society annual meeting*, volume 50, pages 904–908, 2006.

[9] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. Foundations for an empirically determined scale of trust in automated systems. *Int. J. cognitive ergonomics*, 4(1):53–71, 2000.

[10] Vivian Lai et al. Towards a science of human-AI decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471*, 2021.

[11] Han Liu, Vivian Lai, and Chenhao Tan. Understanding the effect of out-of-distribution examples and interactive explanations on human-AI decision making. *Proc. ACM Human-Computer Interaction*, 5(CSCW2):1–45, 2021.

[12] Arnold Lund. Measuring usability with the use questionnaire. *Usability and User Experience Newsletter of the STC Usability SIG*, 8, 01 2001.

[13] James A Russell, Anna Weiss, and Gerald A Mendelsohn. Affect grid: a single-item scale of pleasure and arousal. *J. personality and social psychology*, 57(3):493, 1989.

[14] Jeff Sauro and Joseph Dumas. Comparison of three one-question, post-task usability questionnaires. In *Proc. SIGCHI human factors in computing systems*, pages 1599–1608, 2009.

[15] Dakuo Wang et al. Autods: Towards human-centered automation of data science. In *Proc. CHI Conf. Human Factors in Computing Systems*, pages 1–12, 2021.

[16] David Watson and Lee Anna Clark. The panas-x: Manual for the positive and negative affect schedule-expanded form. 1994.

[17] Justin D Weisz et al. Better together? An evaluation of AI-supported code translation. In *27th Int. Conf. Intelligent User Interfaces*, pages 369–391, 2022.

[18] Julia K. Winkler et al. Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. *JAMA Dermatology*, 155(10):1135–1141, 10 2019.

[19] Albert Ziegler et al. Productivity assessment of neural code completion. In *Proc. 6th ACM SIGPLAN Int. Symp. Machine Programming*, pages 21–29, 2022.