Explainable AI for End-Users

Sunnie S. Y. Kim
Princeton University
Princeton, NJ, USA

Elizabeth Anne Watkins
Intel Labs
Santa Clara, CA, USA

Olga Russakovsky
Princeton University
Princeton, NJ, USA

Ruth Fong
Princeton University
Princeton, NJ, USA

Andrés Monroy-Hernández
Princeton University
Princeton, NJ, USA

CHI'23 Human-Centered Explainable AI (HCXAI) Workshop, April 28–29, 2023

Abstract
Explainable AI (XAI) suffers from "inmates running the asylum" where (X)AI researchers develop explanations for themselves, leaving out end-users. To understand end-users' XAI needs, uses, and perceptions, we conducted a mixed-methods study with 20 end-users of a real-world AI application [10]. In this work, we reflect on our findings and answer two questions from the workshop call: Q1. How do user characteristics impact needs around explainability? Q2. What user goals should XAI aim to support? We also expand the discussion and answer: Q3. How do user characteristics impact perceptions of AI explanations? We conclude with suggestions for future research.

Author Keywords
Explainable AI (XAI), Interpretability, Human-Centered XAI, Human-AI Interaction, Human-AI Collaboration, XAI for Computer Vision, Local Explanations

Introduction
Despite Miller and colleagues' warning in 2017 [16], explainable AI (XAI) still suffers from "inmates running the asylum" [4]. As pointed out by recent work [2, 16], XAI methods are developed by (X)AI researchers primarily for (X)AI researchers and practitioners to inspect AI systems. *End-users* are left out in current XAI research and development. This gap is critical because end-users may have

different needs and goals that XAI methods should but don't yet support. End-users may also perceive explanations produced by these methods differently.

To understand end-users' XAI needs, uses, and perceptions, we conducted a mixed-methods study with 20 end-users of a real-world AI application [10]. In this work, we reflect on our findings and answer two questions from the workshop call: Q1. How do user characteristics impact needs around explainability? Q2. What user goals should XAI aim to support? We also expand the discussion and answer: Q3. How do user characteristics impact perceptions of AI explanations? We conclude with suggestions for future research.

Study summary

To connect XAI research and development with end-users, we conducted a case study of a real-world context where XAI methods might be deployed [10]. Our research setting was Merlin, an AI-based mobile phone app that identifies birds in user-uploaded photos and audio recordings (Fig. 1). We chose Merlin because it is a widely-used app that allows us to connect with a diverse set of active end-users. Concretely, we interviewed 20 Merlin end-users who span the range from low-to-high AI background (representing both (X)AI consumers and creators) and low-to-high domain background (representing both users who know *less* and *more* about birding than the AI).[1]

With each participant, we conducted an hour-long interview, which included a survey and an interactive feedback session. We studied participants' **XAI needs** through open-ended questions and a survey we developed from the XAI Question Bank [12]. We also studied participants' **XAI uses and perceptions** by mocking up four representative XAI approaches that could potentially be embedded into the app, i.e., heatmap, example, concept, and prototype-based explanations of the app's outputs. Using screen sharing during the Zoom interview, we showed *real* examples of the app's outputs along with *mock-up* explanations we designed (Fig. 1), and asked for participants' feedback.

[1] See [10] for details on how we recruited and classified participants into domain and AI background subgroups.

Q1. How do user characteristics impact needs around explainability?

First, we found that participants' explainability needs varied depending on their **AI background** and **domain interest**. While participants were generally curious about the AI, only those with high-AI background or high-domain interest had needs for AI system details. However, participants unanimously expressed a need for practically useful information that can improve their collaboration with the AI.

According to the survey results, participants wanted to know everything about the AI. They wanted to know about the AI's training data, performance, and inner workings, as well as how the AI produces outputs on specific inputs. Note that the former can be satisfied by providing more information about the AI, whereas the latter requires XAI methods. The picture changed, however, when we tempered self-reported levels of curiosity with interview questions about the effort participants were willing to invest to satisfy that curiosity.

Most participants said they wouldn't go out of their way to gain more information about the AI. Exceptions were participants with high-AI background or high-domain interest. Participants with **high-AI background**, likely because they develop AI systems in their work, were very curious about the AI and were willing to go to the extent of reaching out to the app developers or playing with the data themselves.

Participants with **high-domain interest** were particularly curious about how the AI identifies birds that are difficult for experienced human birders (e.g., mockingbirds, "little brown birds"). In contrast, participants with **low-to-medium AI background** had lower explainability needs. Some even preferred to keep the AI as a black box, saying they "don't want to ruin the mystique."

While participants' needs for AI system details differed based on background and interest, **all participants** expressed a need for practically useful information. This included general information about the AI's capabilities and limitations, display of the AI's confidence on specific outputs, and more detailed outputs. Participants wanted this information to improve their collaboration with the AI, particularly in deciding when to use the AI and accept its outputs.

Q2. What user goals should XAI aim to support?

In our study, we also identified various user goals XAI methods should aim to support. When we showed mock-up explanations to participants, they were excited to use them to achieve various goals. One was to **understand** how the AI produced the specific output, which is the immediate goal local XAI methods are developed to help people achieve. Another goal was to **calibrate trust** in the AI, which is also a well-known use of XAI [6, 15, 21, 26, 28].

There were also goals newer to the XAI literature. For example, participants wanted to **learn** from the AI via explanations to perform the task (bird identification) better on their own. Participants also wanted actionable feedback on their own behavior that would help them **supply better inputs** to the AI, helping it be more accurate. Finally, participants with high-AI background saw explanations as a medium to **give feedback** to the developers to improve the AI.

The last two goals in particular suggest that participants viewed the AI as a **collaborator**. To improve their collaboration, they wanted explanations to **help them help the AI**, e.g., by supplying better inputs and providing constructive feedback to developers. We found this an intriguing re-purposing of explanations, which are typically provided to help people understand and calibrate their trust in AI. Our findings highlight the broad range of user goals that should be considered in XAI research and development.

Q3. How do user characteristics impact perceptions of AI explanations?

Finally, we expand the discussion in the workshop call and describe how participants' **AI background** impacted their perceptions of explanations. Due to space constraints, we present results for heatmap and concept-based explanations (Fig. 1). See the full paper [10] for details and results for example and prototype-based explanations.

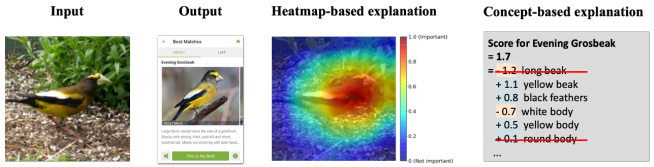

Figure 1: Merlin is an AI-based mobile phone app that identifies birds in user-uploaded photos and audio recordings. In this figure, we show a *real* input-output example and corresponding *mock-up* explanations we designed.

Heatmap-based explanations [7, 19, 22, 24, 25, 27, 29, 3], while deemed intuitive in the AI research community, received mixed reviews from participants. Participants with **high-AI background**, who often use heatmaps in their work, found them intuitive and helpful for representing information. However, participants with **low-AI background** expressed a strong dislike. One participant remarked, *"I*

hate those things [...] They are simply not intuitive." Another participant didn't like them as explanations because *"heatmaps feel like they should be related to weather,"* revealing individual differences in perception.

Similarly, opinions diverged for concept-based explanations [11, 20, 30]. Some participants with **low-AI background** found them confusing. One participant said, *"stuff like this would go right over my head."* Contrarily, participants with **higher-AI background** wanted even more numbers, concepts, and other details about how the AI makes its predictions. Still, they acknowledged that such explanations might overwhelm users who are less familiar with AI.

Discussion & Suggestions for future research

In summary, we observed a creator-consumer gap in XAI [5]. XAI creators I typically assume that consumers are curious about AI system details (e.g., how the AI works and produces outputs). However, we found that participants who are not AI experts like XAI creators had low needs for these details. They didn't want to go out of their way to learn system details and wanted explanations for more practical purposes, e.g., for determining when to trust the AI, improving their task skills, and helping the AI perform better. Further, these participants found some of the popular XAI approaches confusing, suggesting that the field is not sufficiently considering and supporting end-users' needs.

To close this creator-consumer gap and develop XAI for end-users, we make four suggestions for future research.

(1) Design explanations with end-users. In our study, participants exposed blind spots in existing XAI methods and proposed solutions. For example, they pointed out that the concepts used in concept-based explanations were disconnected from birders' language. The shown concepts (e.g., white body, long wings) were too generic compared to birders' field mark terms (e.g., wingbar, supercilium). To solve this disconnect, they suggested developing the bank of concepts with domain experts and end-users like themselves, and offered to contribute their expertise. This example highlights the need for and benefit of end-users' participation in the XAI design process and calls for more participatory approaches [17] to XAI.

(2) Design explanations that answer "why" not just "what." Some participants were unsatisfied with existing XAI approaches that only explain "what" features the AI system was using to produce its output, e.g., heatmap explanations that highlight "what" image regions were important but don't explain "why" those regions were important. They expressed a desire for explanations that answer "why" question so that they can gain a deeper understanding of how the AI works and produces outputs.

(3) Design explanations that use multiple forms and modalities. Participants often suggested combining two or more XAI approaches to produce more informative explanations. They also questioned why Merlin's identification features and our explanation mock-ups were not multimodal, when human birders combine evidence from as many sources as possible (e.g., photo, sound, location) for more accurate bird identification. Expanding the design space of explanations will better satisfy end-users' needs.

(4) Evaluate explanations rigorously. Explanations sometimes have (unintended) negative effects. Recent works have revealed that explanations can engender over-trust in AI or give misleading understandings [9, 5, 1, 8, 13, 14, 18, 23]. Our study participants were also concerned about the faithfulness and potential negative effects of explanations. To preemptively, not reactively, address these issues, it is crucial to rigorously evaluate XAI methods throughout their development process.

Funding acknowledgments

This was supported by the Princeton SEAS Howard B. Wentz, Jr. Junior Faculty Award (OR), Princeton SEAS Project X Fund (RF, OR), Princeton Center for Information Technology Policy (EW), Open Philanthropy (RF, OR), NSF Graduate Research Fellowship (SK), and NSF Grants No. 1763642 and 2145198 (OR). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

REFERENCES

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity Checks for Saliency Maps. In *Neural Information Processing Systems (NeurIPS)*.

[2] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. 2020. Explainable Machine Learning in Deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 648–657. DOI: http://dx.doi.org/10.1145/3351095.3375624

[3] Wieland Brendel and Matthias Bethge. 2019. Approximating CNNs with Bag-of-local-Features Models Works Surprisingly Well on ImageNet. In *International Conference on Learning Representations (ICLR)*.

[4] Alan Cooper. 2004. *The Inmates Are Running the Asylum: Why High Tech Products Drive Us Crazy and How to Restore the Sanity (2nd Edition)*. Pearson Higher Education.

[5] Upol Ehsan, Samir Passi, Q. Vera Liao, Larry Chan, I-Hsiang Lee, Michael J. Muller, and Mark O. Riedl. 2021. The Who in Explainable AI: How AI Background Shapes Perceptions of AI Explanations. *CoRR* abs/2107.13509 (2021). https://arxiv.org/abs/2107.13509

[6] Andrea Ferrario and Michele Loi. 2022. How Explainability Contributes to Trust in AI. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1457–1466. DOI: http://dx.doi.org/10.1145/3531146.3533202

[7] Ruth Fong and Andrea Vedaldi. 2017. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In *International Conference on Computer Vision (ICCV)*.

[8] Adrian Hoffmann, Claudio Fanconi, Rahul Rade, and Jonas Kohler. 2021. This Looks Like That... Does it? Shortcomings of Latent Space Prototype Interpretability in Deep Networks. In *International Conference on Machine Learning (ICML) Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI*.

[9] Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, and Olga Russakovsky. 2022. HIVE: Evaluating the Human Interpretability of Visual Explanations. In *European Conference on Computer Vision (ECCV)*.

[10] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. *ACM CHI Conference on Human Factors in Computing Systems (CHI)* (2023).

[11] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept Bottleneck Models. In *International Conference on Machine Learning (ICML)*.

[12] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. *Questioning the AI: Informing Design Practices for Explainable AI User Experiences*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376590

[13] Zachary C. Lipton. 2018. The Mythos of Model Interpretability: In Machine Learning, the Concept of Interpretability is Both Important and Slippery. *Queue* (2018).

[14] Andrei Margeloiu, Matthew Ashman, Umang Bhatt, Yanzhi Chen, Mateja Jamnik, and Adrian Weller. 2021. Do Concept Bottleneck Models Learn as Intended?. In *International Conference on Learning Representations (ICLR) Workshop on Responsible AI*.

[15] Tim Miller. 2022. Are we measuring trust correctly in explainability, interpretability, and transparency research? (2022). DOI: http://dx.doi.org/10.48550/ARXIV.2209.00651

[16] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. *CoRR* abs/1712.00547 (2017). http://arxiv.org/abs/1712.00547

[17] Michael J. Muller. 2002. *Participatory Design: The Third Space in HCI*. L. Erlbaum Associates Inc., USA, 1051–1068.

[18] Giang Nguyen, Daeyoung Kim, and Anh Nguyen. 2021. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. In *Neural Information Processing Systems (NeurIPS)*.

[19] Vitali Petsiuk, Abir Das, and Kate Saenko. 2018. RISE: Randomized Input Sampling for Explanation of Black-box Models. In *British Machine Vision Conference (BMVC)*.

[20] Vikram V. Ramaswamy, Sunnie S. Y. Kim, Nicole Meister, Ruth Fong, and Olga Russakovsky. 2022. ELUDE: Generating interpretable explanations via a decomposition into labelled and unlabelled features. (2022). DOI: http://dx.doi.org/10.48550/ARXIV.2206.07690

[21] Nicolas Scharowski, Sebastian A. C. Perrig, Nick von Felten, and Florian Brühlmann. 2022. Trust and Reliance in XAI – Distinguishing Between Attitudinal and Behavioral Measures. (2022). DOI: http://dx.doi.org/10.48550/ARXIV.2203.12318

[22] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In *International Conference on Computer Vision (ICCV)*.

[23] Hua Shen and Ting-Hao Kenneth Huang. 2020. How Useful Are the Machine-Generated Interpretations to General Users? A Human Evaluation on Guessing the Incorrectly Predicted Labels. In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.

[24] Vivswan Shitole, Fuxin Li, Minsuk Kahng, Prasad Tadepalli, and Alan Fern. 2021. One Explanation is Not Enough: Structured Attention Graphs for Image Classification. In *Neural Information Processing Systems (NeurIPS)*.

[25] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *International Conference on Learning Representations (ICLR) Workshops*.

[26] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. DOI:http://dx.doi.org/10.1145/3290605.3300509

[27] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision (ECCV)*.

[28] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. DOI: http://dx.doi.org/10.1145/3351095.3372852

[29] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning Deep Features for Discriminative Localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[30] Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. 2018. Interpretable basis decomposition for visual explanation. In *European Conference on Computer Vision (ECCV)*.