Establishing Appropriate Trust in AI through Transparency and Explainability

Sunnie S. Y. Kim
sunniesuhyoung@princeton.edu
Princeton University
Princeton, New Jersey, USA

ABSTRACT

As AI systems are increasingly transforming our society, it is critical to support relevant stakeholders to have appropriate understanding and trust in these systems. My dissertation research explores how providing transparency and explainability for AI systems can help with this goal. I begin with human-centered evaluations of current AI explanation techniques, focusing on their usefulness for people in understanding model behavior and calibrating trust. Next, I identify what explainability needs actual end-users have and what factors influence their trust through an in-depth case study of a real-world AI application. Finally, I describe two studies, one ongoing and one proposed, that investigate transparency and explainability approaches for Generative AI, such as large language models, to enable safe and successful interactions with this new and powerful technology. My dissertation contributes to both HCI and AI fields by elucidating mechanisms and factors of trust in AI and detailing design considerations for AI transparency and explainability approaches.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; • **Computing methodologies → Artificial intelligence**.

KEYWORDS

AI transparency and explainability, Explainable AI, Trust and reliance, Human-AI collaboration

ACM Reference Format:
Sunnie S. Y. Kim. 2024. Establishing Appropriate Trust in AI through Transparency and Explainability. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3613905.3638184

FOREWORD

I am currently a fourth-year PhD student in the Computer Science department at Princeton University, advised by Professor Olga Russakovsky. I work on AI transparency and explainability to help people better understand and interact with AI systems. My research has been published in both HCI and AI venues (e.g., CHI, FAccT, CVPR, ECCV) and is supported by the NSF Graduate Research Fellowship. I have completed all required coursework and successfully passed the program's general exam in my second year. The expected completion date of my PhD studies is Spring 2025. Upon graduation, I plan to seek research positions where I can continue investigating factors that lead to appropriate trust in AI and advocating for the importance of transparency and explainability.

1 CONTEXT AND MOTIVATION

Appropriate trust is key to safe and successful interactions with AI systems. Despite the rapid growth of technology, AI systems still frequently and unexpectedly fail for various reasons. Users who know when and how much to trust an AI system are likely to effectively use the system and achieve their goals. In contrast, users with unwarranted trust (trusting when the AI is untrustworthy) and unwarranted distrust (distrusting when the AI is trustworthy) are likely to have low-quality interactions, even unsafe ones when they are using the AI in high-stakes settings [1, 17, 36, 49].

Inherently linked to trust in AI are transparency and explainability. Based on information provided by various transparency and explainability approaches, people form an understanding of an AI system's capabilities and limitations, how it works, and how it produced a specific output—and this understanding forms the basis of their trust. Hence, AI transparency and explainability approaches are often viewed as trust calibration methods. Transparency approaches include providing model cards [37], model internals [40], performance measures [23, 27, 43, 51, 52], and (un)certainty information [39, 53]. Explainability approaches include providing local explanations about specific model outputs [13, 16, 24, 50, 54] and global explanations about what the model has learned [2, 3, 15] and how it recognizes a specific class [25, 42, 55]. To date, hundreds of approaches have been proposed, varying in explanation process (e.g., feature attribution [13, 44, 54], counterfactual examples [16, 46]) and form (e.g., heatmap-based [14, 44, 54], part-based [8, 42, 55]).

In contrast to the rapid development of approaches, understandings of when and how AI transparency and explainability lead to (or don't lead to) appropriate understanding and trust fall far behind. I attribute this to the lack of human-centered and context-driven research, and in my dissertation, foreground the people who use AI systems and their needs, goals, and contexts. More concretely, my dissertation makes three key contributions. First, it presents insights from human-centered studies of AI transparency and explainability. Much of existing research focuses on the *technology*, i.e., on developing new techniques and evaluating their technical properties, rather than the *people* who will use or be affected them, or the *context* where they will be deployed. In contrast, my dissertation joins the growing body of work that takes a human-centered

CHI EA '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0331-7/24/05.
https://doi.org/10.1145/3613905.3638184

perspective on AI transparency and explainability [9–11, 28–30]. Through two human evaluation studies [20, 41], it sheds light on how useful proposed approaches are to people, in particular for understanding model behavior and calibrating trust. Then through an in-depth case study of a real-world AI application [21], it surfaces actual end-users' explainability needs and goals that are different from those prioritized in current research.

Second, this dissertation provides a more holistic and nuanced understanding of trust in AI. While trust in AI research is fast-growing, there is a lack of empirical studies that approach trust holistically or capture contextual aspects of trust. Most studies are controlled lab experiments that investigate one specific aspect of trust with hypothetical end-users. While they provide valuable insights, studies in real-world contexts are crucial because actual end-users' trust relationships with AI may be different from what researchers anticipate. Through a contextually-grounded study [22], this dissertation elaborates on multiple aspects of actual end-users' trust in AI (e.g., trustworthiness perceptions, trust attitudes, trust-related behaviors) and human, AI, and context-related factors that influence it (see Tab. 1), expanding the field's understanding of mechanisms and factors of trust.

Finally, this dissertation contributes to the development of transparency and explainability approaches for Generative AI, arguably one of the most influential technologies in the current era. For immediate insights, it investigates a transparency approach that can be implemented forthwith for large language models (LLMs) — uncertainty expression through natural language — and examines its effect on user reliance and trust. This study will provide actionable suggestions for when and how LLMs should express uncertainty. For the longer-term, this dissertation proposes a fundamental study of LLMs' self-generated explanations (i.e., explanations of their own answers and behaviors) that builds on explanations research in psychology and cognitive sciences. The anticipated outcome is a principled framework for explainability for LLMs that will guide the field's future research.

2 RESEARCH QUESTIONS

In summary, my dissertation aims to elucidate mechanisms and factors of trust in AI, and develop AI transparency and explainability approaches that help people form appropriate understanding and trust in AI. To this end, it explores the following research questions grouped into three themes.

(1) Human-centered evaluation of AI explanation techniques
  - RQ1-1: Do current AI explanation techniques help people calibrate their trust? [20]
  - RQ1-2: What are the challenges of using current AI explanation techniques in practice? [41]

(2) Contextually-grounded study of explainability needs and trust in AI
  - RQ2-1: What AI explainability needs do end-users have, and how do they perceive current explanation approaches? [21]
  - RQ2-2: What factors influence end-users' trust in AI? [22]

(3) Investigation of transparency and explainability approaches for Generative AI
  - RQ3-1: How do LLMs' natural language expressions of uncertainty affect user reliance and trust? [*ongoing*]
  - RQ3-2: How do people perceive and act upon LLMs' self-explanations? [*proposed*]

3 RESEARCH METHODS

I use a variety of research methodologies, from running computational experiments with large-scale AI models and datasets to conducting quantitative and qualitative user studies. I choose the specific method based on the purpose of the study. For [20, 41], I implemented, trained, and analyzed numerous AI models and explanation techniques. For [20, 41], I also designed user studies, developed study UIs with HTML and Javascript, and conducted experiments on Amazon Mechanical Turk (MTurk). For [21, 22], I conducted semi-structured interviews and analyzed the gathered qualitative data with thematic and abductive coding.

4 FINDINGS TO DATE

4.1 RQ1-1: Do current AI explanation techniques help people calibrate their trust? [20]

The first piece of my dissertation is HIVE [20], a novel evaluation framework for AI explanation techniques. HIVE allows for falsifiable hypothesis testing, cross-method comparison, and human-centered evaluation of AI explanations' ability to help people calibrate their trust in model predictions. My collaborators and I developed it and used it to evaluate four popular techniques (GradCAM [44], BagNet [5], ProtoPNet [7], ProtoTree [38]) with 950 participants recruited from MTurk. Notably, we found that participants struggled to distinguish correct and incorrect model predictions based on explanations. Participants also relied more on model predictions, even incorrect ones, when provided explanations. In other words, popular AI explanation techniques engendered over-trust and overreliance on AI. This finding issued a warning to the field that AI explanation techniques, even when developed with the best of intentions, can have unintended negative effects in human-AI interaction. The full paper was published at ECCV 2022 [20]. Shorter versions also appeared at CHI 2022's Human-Centered Explainable AI workshop and CVPR 2022's Explainable AI for Computer Vision and Women in Computer Vision workshops.

4.2 RQ1-2: What are the challenges of using current AI explanation techniques in practice? [41]

Continuing the line of work on human-centered evaluation of AI explanations, we next examined factors that affect AI explanations' usefulness in practice. We focused on a class of techniques called concept-based explanations that explain model components and predictions with semantic concepts. They are a particularly promising approach for bridging the gap between complex AI models and human understanding, as they explain AI in units that are intuitive to humans (i.e., semantic concepts). Through an in-depth analysis of four representative techniques (NetDissect [2], TCAV [18], Concept Bottleneck [26], IBD [55]) on multiple datasets (ADE20k [56, 57], Pascal [12], CUB-200-2011 [48]), we identified three commonly overlooked factors that have a huge effect on explanation quality:

Table 1: Factors of trust in AI identified in our recent work [22]. Different from most existing work, we explored multiple aspects of trust in AI in a real-world context with actual end-users and identified trust-influencing factors in a bottom-up manner. We organized the factors based on whether they are related to the human trustor, the AI trustee, or the context.

Human-related factors	AI-related factors	Context-related factors
Domain knowledge	Ability	Task difficulty
Ability to assess the AI's outputs	Integrity	Perceived risks and benefits
Ability to assess the AI's ability	Benevolence	Situational characteristics
Ability to use the AI	Popularity	Domain's reputation
	Familiarity	Developers' reputation
	Ease of use	

(1) the choice of the probe dataset on which explanations are generated, (2) the learnability of concepts in the probe dataset, and (3) the number of concepts used in explanations. We also made immediate suggestions for each factor to improve the usefulness of concept-based explanations. Overall, this work highlights the importance of vetting intuitions when developing and using AI explanation techniques, and equips researchers and practitioners with tools to do so. The full paper was published at CVPR 2023 [41].

4.3 RQ2-1: What AI explainability needs do end-users have, and how do they perceive current explanation approaches? [21]

The aforementioned works [20, 41] are among the first human-centered evaluations of AI explanation techniques. Still, their evaluation setups lacked *context* because the study participants, i.e., MTurk workers, were not actual end-users of the AI models being explained. Hence, in our next project [21], we interviewed 20 end-users of a widely-used AI application, the Merlin app for bird identification [45], to study what explainability needs end-users have in a real-world context and how they perceive different explanation approaches (heatmap, example, concept, or prototype-based). Intriguingly, we found that participants desired AI explanations for various purposes beyond understanding the AI system, such as learning domain knowledge from the system and giving feedback to developers, expanding the field's understanding of explainability needs. We also found that participants' perceptions of different explanation approaches vary with respect to their domain and AI knowledge base. These findings provide insights into the nuances of real-world human-AI interactions and highlight that human-centered and contextually-grounded research is necessary to develop effective AI transparency and explainability approaches. The full paper was published at CHI 2023 [21] and received a best paper honorable mention award. Shorter versions also appeared at NeurIPS 2022's Human-Centered AI workshop and CHI 2023's Human-Centered Explainable AI workshop.

4.4 RQ2-2: What factors influence end-users' trust in AI? [22]

In the same interviews, we also inquired about participants' trust in the AI application from many angles with questions about their typical use of the app, as well as whether they would use the app in hypothetical higher-risk scenarios. We analyzed this data in a separate paper [22]. Different from most prior work, which investigates one aspect of trust, we analyzed multiple aspects of trust based on the seminal trust model by Mayer et al. [34] that delineates trust from its antecedents, context, and products. Our holistic approach to trust revealed a comprehensive picture of end-users' trust relationships with AI that cannot be gained by studying only one aspect of trust. Notably, we found a discrepancy between participants' *general* trustworthiness perceptions and trust attitudes, and *instance-specific* trust-related behaviors, adding nuances to existing understandings of trust in AI. Our bottom-up study approach also allowed us to identify a wide range of trust-influencing factors, organized in Table 1 based on whether they are related to the human trustor, the AI trustee, or the context. This work deepens the field's understanding of mechanisms and factors of trust, and yield insights into how readily existing theories of trust (e.g., Mayer et al.'s trust model [34]) can be operationalized for empirical research. The full paper was published at FAccT 2023 [22]. A shorter version appeared at CHI 2023's Trust and Reliance in AI-assisted Tasks workshop.

5 EXPECTED NEXT STEPS

5.1 RQ3-1: How do LLMs' natural language expressions of uncertainty affect user reliance and trust? [19]

Based on the insights from my prior work, I am currently investigating transparency and explainability approaches for Generative AI that is having a rapidly growing impact on our society. As a first step, I have been examining the impact of LLMs' uncertainty expression, a form of transparency [4], on user reliance and trust. This work began in Summer 2023 during my internship at Microsoft Research in the FATE (Fairness, Accountability, Transparency, and Ethics in AI) group. The goal of this work is to understand whether LLMs' natural language expressions of uncertainty can reduce over-reliance and over-trust, a well-known pitfall that has been shown to reduce task performance and worsen user experience [6, 47]. We explore this question in the context of users seeking medical information with LLM-infused search engines such as Copilot in Bing as they are already used by millions of people, and because search is a domain where the factual correctness of AI answers is fundamental. Over the past months, we carefully designed and ran a large-scale, pre-registered, human-subject experiment (N=404). Our findings suggest that using natural language expressions of uncertainty can be an effective approach for reducing overreliance and over-trust on LLMs, but that the precise language used matters: expressions from a first-person perspective (e.g., "*I'm not sure, but...*") were more effective than expressions from a general perspective (e.g., "*There is uncertainty, but...*") in our experiment. We anticipate our findings will inform the design of a wide range of LLM-infused applications. The full paper will be published at FAccT 2024 [19].

5.2 RQ3-2: How do users perceive and act upon explanations from LLMs? [*proposed*]

The final piece of my dissertation will be an investigation of explanations from LLMs. Different from other types of AI models (e.g., classification models), LLMs can and often provide explanations for their answers and behaviors (e.g., "*The answer to your question is X because...*", "*I cannot handle your request because...*"), even when explanations were not requested. I am particularly interested in these explanations because they can be highly convincing and fluent, while lacking faithfulness, relevance, and other desirable properties of explanations. They are also likely already impacting millions of users, but very little is known about how users perceive and act upon them. To tackle this problem, I will first develop a taxonomy of properties for LLMs' self-generated explanations, building on the rich literature on explanations from psychology and cognitive sciences [31, 32, 35]. With this taxonomy, I will then investigate how different properties affect user satisfaction and trust in explanations from LLMs. For example, prior research in psychology has found that people prefer selective explanations [33]. I am curious if this finding will still hold for LLM explanations and whether there are other more important properties. The anticipated outcome of this work is a principled framework for explainability for LLMs, that can guide the field's future research and design of LLM explanations.

6 RESEARCH COMMUNITY ACTIVITIES

Beyond my research work, I am passionate about building and connecting research communities. This past year, I helped build a community for Explainable AI (XAI) researchers and practitioners by co-managing the ExplainableAIWorld slack group (380+ people) and the @XAI_Research twitter account (1600+ followers). I also contributed to organizing a talk series for junior researchers to share their work and meet other researchers. In June 2023, I led the organization of the Explainable AI for Computer Vision (XAI4CV) workshop at CVPR 2023. In organizing the workshop, I made active efforts to bridge the HCI and AI research communities by inviting distinguished researchers in both fields as keynote speakers, and creating a new demo track to emphasize the role of communication and interaction in explainability research. The workshop was hugely successful with over 200 people in attendance. Currently, I am part of two workshop organizing committees. One is for another iteration of the XAI4CV workshop at CVPR 2024. Another is for the Human-Centered Explainable AI (HCXAI) workshop at CHI 2024. I hope these workshops further connect the HCI and AI research communities and encourage diverse perspectives and approaches to AI transparency and explainability.

7 CONCLUSION

Appropriate trust is key to safe and successful interactions with AI systems. My dissertation aims to elucidate mechanisms and factors of trust in AI and develop AI transparency and explainability approaches that help people form appropriate understanding and trust. As a first step, I conducted human-centered evaluations of current AI explanation techniques' ability to help people understand model behavior and calibrate their trust. I then conducted an in-depth case study of a real-world AI application and identified actual end-users' explainability needs and trust relationship with AI. As next steps, I described two studies, one ongoing and one proposed, that investigate transparency and explainability approaches for Generative AI that is having a transformative impact on our society. Together, this dissertation contributes to both HCI and AI fields and lays out actionable steps for establishing appropriate understanding and trust in AI.

ACKNOWLEDGMENTS

I am grateful to my PhD adviser and research collaborators for their mentorship and support. I also thank the participants in my research studies for generously sharing their time and experiences. The studies mentioned in this work were supported by the NSF Graduate Research Fellowship and other funding to my PhD adviser and research collaborators: NSF Grants No. 1763642 and 2145198, Princeton SEAS Howard B. Wentz, Jr. Junior Faculty Award, Princeton SEAS Project X Fund, Open Philanthropy, and Princeton Center for Information Technology Policy.

REFERENCES

[1] Nikola Banovic, Zhuoran Yang, Aditya Ramesh, and Alice Liu. 2023. Being Trustworthy is Not Enough: How Untrustworthy Artificial Intelligence (AI) Can Deceive the End-Users and Gain Their Trust. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 27 (apr 2023), 17 pages. https://doi.org/10.1145/3579460
[2] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network dissection: Quantifying interpretability of deep visual representations. In *CVPR*.
[3] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. 2019. Seeing What a GAN Cannot Generate. In *International Conference on Computer Vision (ICCV)*.
[4] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. 2021. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 401–413. https://doi.org/10.1145/3461702.3462571
[5] Wieland Brendel and Matthias Bethge. 2018. Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet. In *ICLR*.
[6] 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. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (apr 2021), 21 pages. https://doi.org/10.1145/3449287
[7] Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. 2019. This looks like that: Deep learning for interpretable image recognition. In *NeurIPS*.
[8] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. 2019. This Looks Like That: Deep Learning for Interpretable Image Recognition. In *Neural Information Processing Systems (NeurIPS)*.
[9] Upol Ehsan and Mark O. Riedl. 2020. Human-centered Explainable AI: Towards a Reflective Sociotechnical Approach. *CoRR* abs/2002.01092 (2020). arXiv:2002.01092 https://arxiv.org/abs/2002.01092
[10] Upol Ehsan, Philipp Wintersberger, Q. Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O. Riedl. 2021. Operationalizing Human-Centered Perspectives in Explainable AI. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan)

(CHI EA '21). Association for Computing Machinery, New York, NY, USA, Article 94, 6 pages. https://doi.org/10.1145/3411763.3441342

[11] Upol Ehsan, Philipp Wintersberger, Q. Vera Liao, Elizabeth Anne Watkins, Carina Manger, Hal Daumé III, Andreas Riener, and Mark O Riedl. 2022. Human-Centered Explainable AI (HCXAI): Beyond Opening the Black-Box of AI. 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 109, 7 pages. https://doi.org/10.1145/3491101.3503727

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2010. The Pascal Visual Object Classes (VOC) Challenge. *IJCV* (2010).

[13] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. 2019. Understanding Deep Networks via Extremal Perturbations and Smooth Masks. In *International Conference on Computer Vision (ICCV)*.

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

[15] Ruth Fong and Andrea Vedaldi. 2018. Net2Vec: Quantifying and Explaining how Concepts are Encoded by Filters in Deep Neural Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[16] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Counterfactual Visual Explanations. In *International Conference on Machine Learning (ICML)*.

[17] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 624–635. https://doi.org/10.1145/3442188.3445923

[18] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *ICML*.

[19] Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, and Jennifer Wortman Vaughan. 2024. "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In *To appear in ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

[20] 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)*.

[21] 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. In *ACM Conference on Human Factors in Computing Systems (CHI)*.

[22] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. Humans, AI, and Context: Understanding End-Users' Trust in a Real-World Computer Vision Application. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*.

[23] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-User Expectations of AI Systems. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300641

[24] Pang Wei Koh and Percy Liang. 2017. Understanding Black-box Predictions via Influence Functions. In *International Conference on Machine Learning (ICML)*.

[25] 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)*.

[26] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *ICML*.

[27] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 29–38. https://doi.org/10.1145/3287560.3287590

[28] 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

[29] Q. Vera Liao and Kush R. Varshney. 2021. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *CoRR* abs/2110.10790 (2021). arXiv:2110.10790 https://arxiv.org/abs/2110.10790

[30] Q. Vera Liao, Yunfeng Zhang, Ronny Luss, Finale Doshi-Velez, and Amit Dhurandhar. 2022. Connecting Algorithmic Research and Usage Contexts: A Perspective of Contextualized Evaluation for Explainable AI. https://doi.org/10.48550/ARXIV.2206.10847

[31] Emily G. Liquin and Tania Lombrozo. 2020. A functional approach to explanation-seeking curiosity. *Cognitive Psychology* 119 (2020), 101276. https://doi.org/10.1016/j.cogpsych.2020.101276

[32] Emily G. Liquin and Tania Lombrozo. 2022. Motivated to learn: An account of explanatory satisfaction. *Cognitive Psychology* 132 (2022), 101453. https://doi.org/10.1016/j.cogpsych.2021.101453

[33] Tania Lombrozo and Emily G. Liquin. 2023. Explanation Is Effective Because It Is Selective. *Current Directions in Psychological Science* 32, 3 (2023), 212–219. https://doi.org/10.1177/09637214231156106 arXiv:https://doi.org/10.1177/09637214231156106

[34] Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An Integrative Model of Organizational Trust. *The Academy of Management Review* 20, 3 (1995), 709–734. http://www.jstor.org/stable/258792

[35] Tim Miller. 2017. Explanation in Artificial Intelligence: Insights from the Social Sciences. *CoRR* abs/1706.07269 (2017). arXiv:1706.07269 http://arxiv.org/abs/1706.07269

[36] Tim Miller. 2022. Are we measuring trust correctly in explainability, interpretability, and transparency research? https://doi.org/10.48550/ARXIV.2209.00651

[37] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 220–229. https://doi.org/10.1145/3287560.3287596

[38] Meike Nauta, Ron van Bree, and Christin Seifert. 2021. Neural Prototype Trees for Interpretable Fine-Grained Image Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[39] 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)*.

[40] 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* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 237, 52 pages. https://doi.org/10.1145/3411764.3445315

[41] Vikram V. Ramaswamy, Sunnie S. Y. Kim, Ruth Fong, and Olga Russakovsky. 2023. Overlooked Factors in Concept-based Explanations: Dataset Choice, Concept Learnability, and Human Capability. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[42] 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. https://doi.org/10.48550/ARXIV.2206.07690

[43] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I Can Do Better than Your AI: Expertise and Explanations. In *IUI*.

[44] 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 *ICCV*.

[45] The Cornell Lab of Ornithology. [n. d.]. *Merlin Bird ID.* https://merlin.allaboutbirds.org/

[46] Simon Vandenhende, Dhruv Mahajan, Filip Radenovic, and Deepti Ghadiyaram. 2022. Making Heads or Tails: Towards Semantically Consistent Visual Counterfactuals. In *European Conference on Computer Vision (ECCV)*.

[47] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 129 (apr 2023), 38 pages. https://doi.org/10.1145/3579605

[48] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. *The Caltech-UCSD Birds-200-2011 dataset.* Technical Report CNS-TR-2011-001. California Institute of Technology.

[49] Magdalena Wischnewski, Nicole Krämer, and Emmanuel Müller. [n. d.]. Measuring and Understanding Trust Calibrations for Automated Systems: A Survey of the State-Of-The-Art and Future Directions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3544548.3581197

[50] Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep K Ravikumar. 2018. Representer Point Selection for Explaining Deep Neural Networks. In *Neural Information Processing Systems (NeurIPS)*.

[51] 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* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300509

[52] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Jianlong Zhou, and Fang Chen. 2019. Do I Trust My Machine Teammate? An Investigation from Perception to Decision. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 460–468. https://doi.org/10.1145/3301275.3302277

[53] 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 (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852

[54] 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)*.

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

[56] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ADE20k dataset. In *CVPR*.

[57] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2019. Semantic understanding of scenes through the ADE20k dataset. *IJCV* (2019).