AI + HCI for Trustworthy and Appropriately Trusted AI Sunnie S. Y. Kim (Princeton)

As AI systems are increasingly transforming our society, we need to build systems that are trustworthy and appropriately trusted by users. In my research, I tackle both the technical and human sides of this problem by integrating perspectives and methodologies from AI and HCI.

First, to build trustworthy AI systems, I worked on surfacing and correcting biases of AI systems so they don't lead to discrimination. Specifically, I worked on mitigating gender bias [CVPR21] and contextual bias [ReScience21] in visual recognition models by making interventions at both the data and algorithm levels. [CVPR21] was particularly impactful in the field by identifying new challenges and considerations for using generative models to create fairer datasets.

In tandem, I worked on providing explainability for AI systems, another key pillar of trustworthiness. I developed novel AI explanation methods to help users understand and make informed decisions about AI outputs [CVPRW22, CVPRW24]. Then to bridge the gap between research and real-world implementation, I conducted an in-depth analysis of current methods and identified factors that can limit their practical usefulness [CVPR23]. I also interviewed users of a widely-deployed AI application and uncovered what explainability needs they have and how they perceive current methods, connecting technical research with real users [CHI23 🏅].

Equally important to building trustworthy AI systems is ensuring that these systems are appropriately trusted by users. To this end, I worked on deepening the field's understanding of trust in AI. Unlike most prior work that studied one aspect of trust in a lab setting, in [FAccT23], I studied multiple aspects of trust in a real-world context by interviewing real AI users. This enabled a more holistic and nuanced understanding of trust. Notably, I found that participants who generally perceive the AI system as trustworthy still carefully verify the system's outputs and adopt the system in a case-by-case manner. I also identified factors that influence trust in AI that are new or rarely considered in current AI development.

In parallel, I implemented and evaluated various trust calibration strategies. In [ECCV22], I developed HIVE, an evaluation framework that enables falsifiable hypothesis testing, cross-method comparison, and human-centered evaluation of AI explanation methods. With it, I found that current methods don't reliably support trust calibration: explanations engendered trust, even when the AI system was incorrect. Later in [FAccT24], I conducted a large-scale, pre-registered, human-subject experiment and found that LLMs' expression of uncertainty through natural language can be an effective way to reduce overreliance and over-trust in LLMs. However, perspective mattered: LLMs' uncertainty expressions in first-person (e.g., "I'm not sure, but…") had stronger effects than general perspective (e.g., "There is uncertainty, but…").

In addition to my research work bringing together the fields of AI and HCI, I have organized multiple workshops to connect researchers from these two communities. At the Explainable AI for Computer Vision workshops at CVPR 2023 and 2024, I invited HCI researchers as keynote speakers and created a new demo track to emphasize the role of communication and interaction in explainability research. At the Human-Centered Explainable AI workshop at CHI 2024, I created LLM lecture materials and hands-on activities for HCI researchers to explore explainability in LLMs. I also co-managed the ExplainableAIWorld slack (400+ people) and the @XAI_Research twitter account (1900+ followers) to build a community for Explainable AI researchers. I am excited to continue these efforts going forward as I strongly believe AI and HCI communities must work together to build trustworthy and appropriately trusted AI.

AI + HCI for Trustworthy and Appropriately Trusted AI Sunnie S. Y. Kim (Princeton)

Trustworthy AI: Fairness

[CVPR21] V. V. Ramaswamy, S. S. Y. Kim, O. Russakovsky. Fair Attribute Classification through Latent Space De-biasing. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.*

[ReScience21] S. S. Y. Kim, S. Zhang, N. Meister, O. Russakovsky. [Re] Don't Judge an Object by Its Context: Learning to Overcome Contextual Bias. *ReScience C, 2021.*

Trustworthy AI: Explainability

[CVPRW22] V. V. Ramaswamy, S. S. Y. Kim, N. Meister, R. Fong, O. Russakovsky. ELUDE: Generating Interpretable Explanations via a Decomposition into Labelled and Unlabelled Features. *CVPR Workshop on Explainable AI for Computer Vision, 2022.*

[CVPRW24] G. Nguyen, M. R. Taesiri, S. S. Y. Kim, A. T. Nguyen. Allowing Humans to Interactively Guide Machines Where to Look Does Not Always Improve Human-AI Team's Classification Accuracy. *CVPR Workshop on Explainable AI for Computer Vision, 2024.*

[CVPR23] V. V. Ramaswamy, S. S. Y. Kim, R. Fong, O. Russakovsky. Overlooked Factors in Concept-based Explanations: Dataset Choice, Concept Salience, and Human Capability. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.*

[CHI23 🏅] S. S. Y. Kim, E. A. Watkins, O. Russakovsky, R. Fong, A. Monroy-Hernández. "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. *ACM Conference on Human Factors in Computing Systems (CHI), 2023.* **Best paper honorable mention award.**

Appropriately Trusted AI: Understanding Trust in AI

[FAccT23] S. S. Y. Kim, E. A. Watkins, O. Russakovsky, R. Fong, A. Monroy-Hernández. Humans, Context, and AI: Understanding End-Users' Trust in a Real-World Computer Vision Application. *ACM Conference on Fairness, Accountability, and Transparency (FAccT), 2023.*

Appropriately Trusted AI: Evaluating Trust Calibration Strategies

[ECCV22] S. S. Y. Kim, N. Meister, V. V. Ramaswamy, R. Fong, O. Russakovsky. HIVE: Evaluating the Human Interpretability of Visual Explanations. *European Conference on Computer Vision (ECCV), 2022.*

[FAccT24] S. S. Y. Kim, Q. V. Liao, M. Vorvoreanu, S. Ballard, J. W. Vaughan. "I'm Not Sure, But...": Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. *ACM Conference on Fairness, Accountability, and Transparency (FAccT), 2024.*