Sunnie S. Y. Kim

sunniesuhyoung@gmail.com
https://sunniesuhyoung.github.io

EMPLOYMENT

2025–Now **Apple**
Research scientist in the Human-Centered Machine Intelligence & Responsible AI group

2023 **Microsoft Research**
Research intern in the FATE (Fairness, Accountability, Transparency & Ethics in AI) group

EDUCATION

2020–2025 **Princeton University**
PhD in Computer Science
Dissertation: *Advancing Responsible AI with Human-Centered Evaluation*
Committee: Olga Russakovsky (adviser), Andrés Monroy-Hernández
Jennnifer Wortman Vaughan, Q. Vera Liao, Parastoo Abtahi

2019–2020 **Toyota Technological Institute at Chicago**
Visiting student advised by Greg Shakhnarovich

2014–2018 **Yale University**
Bachelor of Science in Statistics and Data Science
GPA 3.91/4.00, *magna cum laude*, Distinction in the major
Recognition for outstanding dedication to the department
Senior thesis advised by John Lafferty

HONORS, AWARDS & FELLOWSHIPS

2025	FAccT 2025 Doctoral Consortium
2025	CHI 2025 Honorable Mention Award 🏅
2025	CHI 2025 Special Recognition for Outstanding Review (2 for Papers, 1 for LBW)
2024	Georgia Tech Doctoral Consortium on Responsible Computing, AI, and Society
2024	MIT Rising Stars in EECS Recognition ⭐
2024	Siebel Scholars Award ($35,000) ⭐
2024	CHI 2024 Doctoral Consortium
2024	Princeton SEAS Travel Grant Award
2023	CHI 2023 Honorable Mention Award 🏅
2023	SIGCHI Gary Marsden Travel Award
2022–2025	NSF Graduate Research Fellowship ($138,000) ⭐
2022–2023	ML Reproducibility Challenge Outstanding Reviewer Award (×2)
2020–2023	Women in Computer Vision Workshop Travel and Registration Award
2018	Yale Adrian Van Sinderen Book Collecting First Prize ($1,000)
2016	Yale Summer Research Fellowship
2014–2018	Korea Presidential Science Scholarship ($200,000) ⭐

PAPERS

Preprints

Measuring and Mitigating Overreliance is Necessary for Building Human-Compatible AI
Lujain Ibrahim, Katherine M. Collins, Sunnie S. Y. Kim, Anka Reuel, Max Lamparth, Kevin Feng, Lama Ahmad, Prajna Soni, Alia El Kattan, Merlin Stein, Siddharth Swaroop, Ilia Sucholutsky, Andrew Strait, Q. Vera Liao, Umang Bhatt

PersonaTeaming: Exploring How Introducing Personas Can Improve Automated AI Red-Teaming
Wesley Hanwen Deng, Sunnie S. Y. Kim, Akshita Jha, Ken Holstein, Motahhare Eslami, Lauren Wilcox, Leon A Gatys

Conference and Journal Publications (Peer-Reviewed)

2025

Fostering Appropriate Reliance on Large Language Models: The Role of Explanations, Sources, and Inconsistencies
Sunnie S. Y. Kim, Jennifer Wortman Vaughan, Q. Vera Liao, Tania Lombrozo, Olga Russakovsky
*ACM Conference on Human Factors in Computing Systems (**CHI**)* 🏅 ***Honorable Mention Award***
(Featured in Microsoft’s New Future of Work Report and presented at 10+ places through invited and contributed talks)

Portraying Large Language Models as Machines, Tools, or Companions Affects What Mental Capacities Humans Attribute to Them
Allison Chen, Sunnie S. Y. Kim, Amaya Dharmasiri, Olga Russakovsky, Judith E. Fan
*Annual Meeting of the Cognitive Science Society (**CogSci**)*

2024

“I’m Not Sure, But...”: Examining the Impact of Large Language Models’ Uncertainty Expression on User Reliance and Trust
Sunnie S. Y. Kim, Q. Vera Liao, Mihaela Vorvoreanu, Stephanie Ballard, Jennifer Wortman Vaughan
*ACM Conference on Fairness, Accountability, and Transparency (**FAccT**)*
(Featured in Axios, New Scientist, ACM showcase, Microsoft’s New Future of Work Report, and the Human-Centered AI Medium publication as “Good Reads in Human-Centered AI”)

2023

“Help Me Help the AI”: Understanding How Explainability Can Support Human-AI Interaction
Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, Andrés Monroy-Hernández
*ACM Conference on Human Factors in Computing Systems (**CHI**)* 🏅 ***Honorable Mention Award***
(One of the top 10 cited CHI papers in 2023–2024 (as of Dec 2024); Featured in the Human-Centered AI Medium publication as “CHI 2023 Editors’ Choice; Invited for talks at multiple AI and HCI conference workshops)

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

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

2022

HIVE: Evaluating the Human Interpretability of Visual Explanations
Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, Olga Russakovsky
*European Conference on Computer Vision (**ECCV**)*
(Selected as spotlight and invited for talks at multiple AI and HCI conference workshops)

Shallow Neural Networks Trained to Detect Collisions Recover Features of Visual Loom-Selective Neurons
Baohua Zhou, Zifan Li, Sunnie S. Y. Kim, John Lafferty, Damon A. Clark
eLife (Journal for the biomedical and life sciences)

2021 **[Re] Don't Judge an Object by Its Context: Learning to Overcome Contextual Bias**
Sunnie S. Y. Kim, Sharon Zhang, Nicole Meister, Olga Russakovsky
ReScience C (Journal for reproducible replications in computational science)

Fair Attribute Classification through Latent Space De-biasing
Vikram V. Ramaswamy, Sunnie S. Y. Kim, Olga Russakovsky
*IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**)*
(Featured in Coursera's GANs Specialization course and the MIT Press Book *Foundations of Computer Vision*; Invited for talks at multiple AI conference workshops)

Information-Theoretic Segmentation by Inpainting Error Maximization
Pedro Savarese, Sunnie S. Y. Kim, Michael Maire, Gregory Shakhnarovich, David McAllester
*IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR**)*

2020 **Deformable Style Transfer**
Sunnie S. Y. Kim, Nicholas Kolkin, Jason Salavon, Gregory Shakhnarovich
*European Conference on Computer Vision (**ECCV**)*

2019 **Which Grades Are Better, A's and C's, or all B's? Effects of Variability in Grades on Mock College Admission Decisions**
Woo-kyoung Ahn, Sunnie S. Y. Kim, Kristen Kim, Peter K. McNally
Judgment and Decision Making (Journal for the psychology of human judgment and decision making)

Workshop Papers and Extended Abstracts (Lightly Peer-Reviewed)

* indicates equal contribution

2025 **Improving Automated LLM Evaluation by Introducing Personas in LLM Red-Teaming**
Wesley Hanwen Deng, Sunnie S. Y. Kim, Akshita Jha, Ken Holstein, Motahhare Eslami, Lauren Wilcox, Leon A Gatys
***NeurIPS** Workshop on LLM Evaluation*

Portraying Large Language Models as Machines, Tools, or Companions Affects What Mental Capacities Humans Attribute to Them
Allison Chen, Sunnie S. Y. Kim, Amaya Dharmasiri, Olga Russakovsky, Judith E. Fan
***CHI** Extended Abstracts (Late Breaking Work)*

Interactivity x Explainability: Toward Understanding How Interactivity Can Improve Computer Vision Explanations
Indu Panigrahi, Sunnie S. Y. Kim*, Amna Liaqat*, Rohan Jinturkar, Olga Russakovsky, Ruth Fong, Parastoo Abtahi
***CHI** Extended Abstracts (Late Breaking Work)*

2024 **Establishing Appropriate Trust in AI through Transparency and Explainability**
Sunnie S. Y. Kim
***CHI** Extended Abstracts (Doctoral Consortium)*

Human-Centered Explainable AI (HCXAI): Reloading Explainability in the Era of Large Language Models (LLMs)
Upol Ehsan, Elizabeth Anne Watkins, Philipp Wintersberger, Carina Manger, Sunnie S. Y. Kim, Niels van Berkel, Andreas Riener, Mark O. Riedl
***CHI** Extended Abstracts (Workshop Proposal)*

Allowing Humans to Interactively Guide Machines Where to Look Does Not Always Improve Human-AI Team's Classification Accuracy
Giang Nguyen, Mohammad Reza Taesiri, Sunnie S. Y. Kim, Anh Nguyen
CVPR *Workshop on Explainable AI for Computer Vision*

2023 **Explainable AI for End-Users**
Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, Andrés Monroy-Hernández
CHI *Workshop on Human-Centered Explainable AI*

2022 **Closing the Creator-Consumer Gap in XAI: A Call for Participatory XAI Design with End-users**
Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, Andrés Monroy-Hernández
NeurIPS *Workshop on Human-Centered AI*

ELUDE: Generating Interpretable Explanations via a Decomposition into Labelled and Unlabelled Features
Vikram V. Ramaswamy, Sunnie S. Y. Kim, Nicole Meister, Ruth Fong, Olga Russakovsky
CVPR *Workshop on Explainable AI for Computer Vision*

2021 **Cleaning and Structuring the Label Space of the iMet Collection 2020**
Vivien Nguyen*, Sunnie S. Y. Kim*
CVPR *Workshop on Fine-Grained Visual Categorization*

White Papers and Technical Reports (Not Peer-Reviewed)

2025 **AI Adoption Across Mission-Driven Organizations**
Dalia Ali, Muneeb Ahmed, Arfa Khan, Hailan Wang, Sunnie S. Y. Kim, Meet Muchhala, Anne Merkle, Orestis Papakyriakopoulos
TUM Think Tank

2018 **Environmental Performance Index**
Zachary A. Wendling, John W. Emerson, Daniel Esty, Marc Levy, Alex de Sherbinin, ..., Sunnie S. Y. Kim, et al.
World Economic Forum (Environmental Performance Index is a large-scale evaluation of 180 countries' environmental health and ecosystem vitality. As the data team lead, I built the full data pipeline and led the analysis work. The results were presented at the World Economic Forum and covered by international media outlets.)

TALKS

2025 NAVER AI Lab & HCI Group

Yonsei CSI 7110 Topics in Responsible AI Course

Princeton COS 598B Machine Behavior Course

Apple Human-Centered Machine Intelligence & Responsible AI Group

Cornell Information Science Colloquium

Johns Hopkins Computer Science Seminar

Boston University Computing & Data Sciences Colloquium

SNU AI Computing Winter School

2024 Cornell Tech Social Technologies Lab

ECCV 2024 Workshop on Explainable Computer Vision

	Princeton Concepts & Cognition Lab
	MILA Human-Centered AI Reading Group
	IBS Data Science Group
	KAIST Kim Jaechul Graduate School of AI
	NYC Computer Vision Day
2023	Explainable AI Talk Series
	CHI 2023 Workshop on Human-Centered Explainable AI
2022	NeurIPS 2022 Workshop on Human-Centered AI
	CVPR 2022 Workshop on Explainable AI for Computer Vision
2021	CVPR 2021 Workshop on Responsible Computer Vision
	CVPR 2021 Workshop for Women in Computer Vision
2020	Princeton Course COS 429 Computer Vision Course
	Princeton PIXL Talk Series
	Princeton Bias in AI Reading Group

ORGANIZING COMMITTEE

2025	FAccT 2025 (Proceedings Co-Chair)
	CVPR 2025 Workshop on Explainable AI for Computer Vision (Co-Organizer)
	NYC Computer Vision Day 2025 (Event Program Committee)
2024	CVPR 2024 Workshop on Explainable AI for Computer Vision (Co-Organizer)
	CHI 2024 Workshop on Human-Centered Explainable AI (Co-Organizer)
2023	CVPR 2023 Workshop on Explainable AI for Computer Vision (Co-Organizer)
	CVPR 2023 Workshop for Women in Computer Vision (Co-Organizer)
2018	NESS NextGen Data Science Day 2018 (Local Organizing Committee)

PROGRAM COMMITTEE & REVIEWING

* indicates special recognitions for outstanding reviews

Conferences

NeurIPS (2025 Main track & Ethics review)

CVPR (2022, 2023, 2024, 2025), ICCV (2021, 2023), ECCV (2022, 2024)

CHI (2023, 2024, 2025**, 2026 AC of Computational Interaction subcommittee)

FAccT (2023, 2024, 2025), AIES (2024), SaTML (2023)

Workshops & Extended Abstracts

CVPR 2025 Workshop on Explainable AI for Computer Vision

CHI 2025 Late Breaking Work*

CHI 2024 Workshop on Human-Centered Explainable AI

CVPR 2024 Workshop on Explainable AI for Computer Vision

NeurIPS 2023 Workshop on Explainable AI in Action

ICML 2023 Workshop on AI & HCI

CVPR 2023 Workshop on Explainable AI for Computer Vision

CVPR 2023 Workshop for Women in Computer Vision

AAAI 2023 Workshop on Representation Learning for Responsible Human-Centric AI

CVPR 2021 Workshop on Responsible Computer Vision

Challenges

ML Reproducibility Challenge (2020, 2021*, 2022*)

Books

Foundations of Computer Vision (Authors: Antonio Torralba, Phillip Isola, and William T. Freeman)

Handbook of Human-Centered Artificial Intelligence (Editor-in-Chief: Wei Xu)

MENTORING

Research Mentoring

2024–2025 **Allison Chen** (CS PhD student at Princeton. Recipient of the NSF Graduate Research Fellowship)
Understanding How People Attribute Mental Capacities to LLMs (paper published in ***CHI EA***)

2024–2025 **Indu Panigrahi** (CS Master's student at Princeton, incoming CS PhD student at UIUC)
Incorporating Interactivity in AI Explanations (paper published in ***CHI EA***)

2022–2023 **Rohan Jinturkar** (CS undergrad at Princeton. Recipient of the Sigma Xi Book Award for Outstanding Undergraduate Research & Outstanding CS Senior Thesis Prize)
Developing an Interactive, Dialogue-based AI Explanation System for Non-Experts (senior thesis)

2020–2022 **Nicole Meister** (ECE undergrad at Princeton, now EE PhD student at Stanford. Recipient of the NSF Graduate Research Fellowship, Calvin Dodd MacCracken Senior Thesis/Project Award & Sigma Xi Book Award for Outstanding Undergraduate Research)
Evaluating AI Explanations & Mitigating Contextual Bias in Visual Recognition Systems (papers published in ***ECCV*** and ***ReScience C***)

2020–2021 **Sharon Zhang** (Math undergrad at Princeton, now CS PhD student at Stanford. Recipient of the Sigma Xi Book Award for Outstanding Undergraduate Research)
Mitigating Contextual Bias in Visual Recognition Systems (paper published in ***ReScience C***)

Non-Research Mentoring

2022–2023 Princeton Computer Science G1 Mentoring Program

2021–2022 Princeton Computer Science Graduate Applicant Support Program

TEACHING

2021 **Princeton Computer Science 429 Computer Vision**
Graduate Teaching Assistant

Princeton AI4ALL
Instructor

2019–2020 **TTI-Chicago Girls Who Code**
Co-Founder and Instructor

2018 **Yale Statistics and Data Science 365/565 Data Mining and Machine Learning**
Undergraduate Teaching Assistant

2017 **Yale Statistics and Data Science 230/530 Data Exploration and Analysis**
Undergraduate Teaching Assistant

OTHER ACTIVITIES

Community Building

2022–2023 Explainable AI Slack and Twitter Community (Co-Organizer)

2017–2019 Yale Dimensions Organization for Women and Other Minorities in Math (Co-Founder)

Volunteering

ECCV (2024), FAccT (2024), CVPR (2022), ICML (2020), ICLR (2020), NeurIPS (2019–2020)

NSF Safety and Trust in AI-Enabled Systems Workshop (2022)

COVID Translate Project (2020)

Committee

2021 Princeton Computer Science Graduate Admissions Committee

2017–2019 Yale Statistics & Data Science Departmental Student Advisory Committee

Updated September 2025