Yiwei Yang

University of Washington

819 Northeast 67th st
Seattle, WA 98115

yanyiwei.github.io
yanyiwei@uw.edu

Education

10/20 — Present *Seattle, WA*	**University of Washington** Ph.D. in Information Science Research Interests: Interpretability, Fairness in Machine Learning - Surfacing and mitigating model biases with interpretability techniques Advisor: Bill Howe
09/15 – 05/19 *Ann Arbor, MI*	**University of Michigan** B.S. in Computer Science and Engineering

Professional Experience

10/20 — Present *Seattle, WA*	**University of Washington** Graduate Student and Researcher
06/22 – 09/22 *Remote*	**SONY AI** Research Intern (Mentors: William Thong, Alice Xiang)
05/18 – 08/18 *San Jose, CA*	**IBM Research, Almaden** Research Intern (Mentors: Eser Kandogan, Prithviraj Sen, Yunyao Li)

Research Projects

02/22 — Present *UW*	**Mitigating Bias of Vision Models with Concept Regularization** Working on a bias mitigation method by reducing the model's sensitivity to some sensitive concept with adversarial learning *In preparation for submission to ICCV 2023*
06/22 — Present *SONY AI*	**Bias Propagation in Knowledge Distillation for Vision Models** Working on measuring and reducing the bias transferred from large to small models during knowledge distillation

05/22 — 11/22	**Surfacing Gender Bias of Vision-Language Models**
UW	Showed that CLIP-based models tend to objectify women through a series of experiments (e.g. embedding association tests)
	Preprint - https://arxiv.org/abs/2212.11261
	In submission to FAccT 2023
05/21 — 01/22	**Measuring and Mitigating Procedural Fairness**
UW	Worked on a novel fairness metric for procedural fairness by computing the difference of feature attributions between groups
	In preparation for submission to AIES 2023

Publications

Conference Full Papers

C.04   A. Lundgard, **Y. Yang**, M.L. Foster, W.S. Lasecki. Bolt: Instantaneous Crowdsourcing via Just-in-Time Training. *In Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2018). Montreal, Canada.*

C.03   S.W. Lee, Y. Zhang, I. Wong, **Y. Yang**, S. D. O'Keefe, W.S. Lasecki. SketchExpress: Remixing Animations for More Effective Crowd-Powered Prototyping Of Interactive Interfaces. *In Proceedings of the ACM Symposium on User Interface Software and Technology (UIST 2017). Quebec City, Canada.*

C.02   H. Kaur, M. Gordon, **Y. Yang**, J. Teevan, E. Kamar, J. Bigham, W.S. Lasecki. CrowdMask: Using Crowds to Preserve Privacy in Crowd-Powered Systems via Progressive Filtering. *In AAAI Conference on Human Computation Demos (HCOMP 2017), Quebec City, Canada.*

C.01   Y. Chen, S.W. Lee, Y. Xie, **Y. Yang**, W.S. Lasecki, S. Oney. Codeon: On Demand Software Development Assistance. *In Proceedings of the International ACM Conference on Human Factors in Computing Systems (CHI 2017), Denver, USA.*

Workshop/Demo/Posters

P.03   **Y. Yang.**, E. Kandogan, Y. Li, W.S.Lasecki, P.Sen. HEIDL: Learning Linguistic Expressions with Deep Learning and Human-in-the-Loop. *In Proceedings of the Association for Computational Linguistics (ACL 2019). Florence, Italy.*
(**Best Poster** at Michigan AI Symposium, 1/55)

P.02   **Y. Yang.**, E. Kandogan, Y. Li, W.S.Lasecki, P.Sen. A study on Interaction in Human-in-the-Loop Machine Learning for Text Analytics. *Joint Proceedings of the ACM IUI 2019 Workshops co-located with the 24th ACM Conference on Intelligent User Interfaces (ACM IUI 2019), Los Angeles, USA,* March 20, 2019.

P.01 S. W. Lee, **Y. Yang**, S. Yan, Y. Zhang, I. Wong, Z. Yan, M. McGruder, C. M. Homan, W. S. Lasecki. Creating Interactive Behaviors in Early Sketch by Recording and Remixing Crowd Demonstrations. *In AAAI Conference on Human Computation Demos (HCOMP 2016), Austin, TX.*