Yunfeng Zhang

Yorktown Heights, New York • zywind@gmail.com • linkedin.com/in/zywind/

SUMMARY

- AI researcher dedicated to developing fair and interpretable AI algorithms and applications.
- Experienced HCI researcher, routinely conduct usability design and evaluation
- Filed 13 patent applications and published over 40 papers with more than 700 citations.

EXPERIENCE

Research Scientist May 2016 - Present

IBM Research, AI Engineering | T. J. Watson Research Center, New York

Awards

- Outstanding Accomplishment Award for Research Advancements to Conversational Technology, 2020.
- Outstanding Research Accomplishment Award for Trustworthy AI, 2019.

Trustworthy AI

- Research and develop methods to detect machine learning biases (unfair to people of different demographics) and develop algorithms that remediate biases. Developer of the IBM's open-source **AI Fairness 360 Toolkit**.
- Design methods for explaining AI predictions to help users understand and debug models, build trust with users, remediate cognitive biases, and improve decision making. Developer of the IBM's open-source **AI Explainability 360 Toolkit**. This and the above work have received **IBM outstanding research accomplishment award.**
- Research and develop methods to communicate AI uncertainty to users to help them make better decisions.
- Develop an A/B testing framework. Design and conduct online experiments to evaluate and compare algorithms.

AI Model Lifecycle Management and AutoAI

- Developed framework components to manage AI lifecycle, in particular the active learning component that helps users continuously improve their models.
- Helped design and develop IBM **Watson OpenScale**. Developed algorithms that monitor and detect feature drift. Helped design model fairness monitoring components.
- Designed and developed visualization techniques that help users compare models generated by IBM **AutoAI**.

Chatbot Development Framework

- Led the design and development of an AI-driven chatbot development framework that combines intent classification, NLP, and AI planning together to improve chatbot developer experience. Parts of the framework were adopted by **Watson Assistant** and led to an **IBM outstanding accomplishment award**.

Postdoctoral Researcher June 2015 - May 2016

IBM Research, Cognitive Environments Lab | T. J. Watson Research Center, New York

- Designed and implemented AI-driven multimodal interaction techniques for smart meeting rooms by incorporating gesture, speech, and face recognition techniques.
- Designed and developed CELIO, an application development framework for distributed, multimodal applications.

Research Intern May 2014 - September 2014

IBM Research, Cognitive Environments Lab | T. J. Watson Research Center, New York

- Researched methods to remediate human cognitive biases in AI-assisted human decision making.
- Designed and implemented an online experiment to collect human decisions under risk.

Research Intern May 2013 – December 2013

Palo Alto Research Center | Palo Alto, California

- Developed computational models to simulate and predict how humans detect changes in stochastic environments.

Graduate Research Assistant September 2009 – June 2015
University of Oregon | Eugene, Oregon

- Developed computational models of human cognition and performance in the context of human-computer interaction.
- Conducted human experiments to study human visual search and multitasking behaviors.
- Developed eye tracking algorithms and analysis software.

EDUCATION

University of Oregon | Ph.D. in Computer and Information Science 2015
University of Oregon | Master of Science in Computer and Information Science 2013
Beijing Normal University | Bachelor of Science in Information Science and Technology 2007

PUBLICATIONS and PATENTS

I have published over 40 papers with more than 700 citations and filed 13 patent applications. For more details, check out my Google scholar page. Selected publications:

- Model Agnostic Multilevel Explanations.
- AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias
- One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques
- Joint Optimization of AI Fairness and Utility: A Human-Centered Approach
- Data Augmentation for Discrimination Prevention and Bias Disambiguation

AWARDS

- Annual Conference of the Cognitive Science Society, Computational Modeling Award for Applied Cognition, 2014.
- ACM CHI Conference on Human Factors in Computing Systems, Best Paper award, 2014.
- First place, Green Driver Programming Contest, 2011.
- First place, Fifth Annual UO Eugene Luks Programming Contest, 2011.
- ACM CHI Conference on Human Factors in Computing Systems, Honorable Mention award, 2010.
- International Conference on Cognitive Modeling, Siegel-Wolf Award for Best Applied Paper, 2010.

TECHNICAL SKILLS

- Proficient in Python, Java, and R. Familiar with C++, Scala, and Julia.
- Proficient in user study protocols and methods, including grounded theory, participatory design, A/B tests, etc.
- Proficient in data visualization and analysis techniques such as Matplotlib, ggplot, ANOVA, and regression.
- Proficient in various machine learning and NLP techniques and libraries.
- Proficient in full stack development.
  - Developed backends with Nodejs express, Java/Scala Play, Vert.x, WSGI and ASGI Python frameworks.
  - Developed frontends with Vue and React.
  - Developed databases using PostgreSQL and MongoDB. Familiar with ORMs such as SQLAlchemy.

Last update: Apr. 15, 2021