Xuhui Zhou **Last update: Oct 08, 2023**

Contact

4902 Forbes Ave
Gates Hillman Complex,
Pittsburgh, PA

Homepage: https://xuhuiz.com/
✉ E-mail: xuhuiz@cs.cmu.edu
Tel: 206-306-5850

Research

Facilitate pro-social social agents that interact cooperatively and positively, align with human values, and contribute to the well-being of individuals and society.

Education

Carnegie Mellon University, Pittsburgh, PA — Aug 2022
PhD in Computer Science (Language Technologies)

University of Washington, Seattle, WA — Sep 2019–Jun 2021
M.Sc in Computational Linguistics
Advisor: Noah Smith

Nanjing University, Nanjing, China — Sep 2015–Jun 2019
B.Sc in Statistics, Department of Mathematics
Advisor: Shujian Huang

University of California Berkeley (visiting student), Berkeley, CA — Aug 2017-May 2018

Publications
(* Equal contribution)

SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents
Xuhui Zhou*, Hao Zhu*, Leena Mathur, Ruohong Zhang, Zhengyang Qi, Haofei Yu, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, Maarten Sap
Under review of ICLR, 2024

WebArena: A Realistic Web Environment for Building Autonomous Agents
Shuyan Zhou, Frank F. Xu, Hao Zhu, **Xuhui Zhou**, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, Graham Neubig
Under review of ICLR, 2024

Can LLMs Keep a Secret? Testing Privacy Implications of Language Models via Contextual Integrity Theory
Niloofar Mireshghallah*, Hyunwoo Kim*, **Xuhui Zhou**, Yulia Tsvetkov, Maarten Sap, Reza Shokri, Yejin Choi
Under review of ICLR, 2024

FANTOM: A Benchmark for Analyzing Theory of Mind in Conversations
Hyunwoo Kim, Melanie Sclar, **Xuhui Zhou**, Ronan Le Bras, Gunhee Kim, Yejin Choi, Maarten Sap
EMNLP, 2023

Don't Take This Out of Context! On the Need for Contextual Models and Evaluations for Stylistic Rewriting
Akhila Yerukola, **Xuhui Zhou**, Maarten Sap
EMNLP, 2023

Learning to translate by learning to communicate
C.M. Downey*, **Xuhui Zhou***, Leo Z. Liu, Shane Steinert-Threlkeld
EMNLP MRL, 2023

Clever Hans or Neural Theory of Mind? Stress Testing Social Reasoning in Large Language Models
Natalie Shapira, Mosh Levy, Hossein Seyed Alavi, **Xuhui Zhou**, Yejin Choi, Yoav Goldberg, Maarten Sap, and Vered Shwartz
Preprint, 2023

Cobra Frames: Contextual Reasoning about Effects and Harms of Offensive Statements
Xuhui Zhou, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D. Hwang, Swabha Swayamdipta, Maarten Sap
Findings of ACL, 2023

Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection
Maarten Sap, Swabha Swayamdipta, Laura Vianna, **Xuhui Zhou**, Yejin Choi, Noah A. Smith
NAACL, 2022

Emergent Communication Fine-tuning (EC-FT) for Pretrained Language Models
Shane Steinert-Threlkeld, **Xuhui Zhou**, Zeyu Liu, C. M. Downey
ICLR EmeCom, 2022
🥇 Runner-up Best Paper

Extracting and Inferring Personal Attributes from Dialogue
Zhilin Wang, **Xuhui Zhou**, Rik Koncel-Kedziorski, Alex Marin, Fei Xia
ACL ConvAI, 2022

Challenges in Automated Debiasing for Toxic Language Detection
Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Noah A.Smith, Yejin Choi
EACL, 2021

Linguistically-Informed Transformations (LIT): A Method for Automatically Generating Contrast Sets
Chuanrong Li*, Lin Shengshuo*, Zeyu Liu*, Xinyi Wu*, **Xuhui Zhou***, Shane Steinert-Threlkeld
EMNLP BlackboxNLP, 2020

Multilevel Text Alignment with Cross-Document Attention
Xuhui Zhou, Nikolaos Pappas, Noah A. Smith
EMNLP, 2020

Evaluating Commonsense in Pre-trained Language Models
Xuhui Zhou, Yue Zhang, Leyang Cui, Dandan Huang
AAAI, 2020

RPD: A Distance Function Between Word Embeddings
Xuhui Zhou, Zaixiang Zheng, Shujian Huang
ACL Student Research Workshop, 2020

Service

Origanizing:

- Theory-of-Mind Workshop at ICML 2023
- LTI Student Research Symposium 2023

Program Committee & Reviewing:

Journals & Conferences

- **TLMR** 2023
- **ACL** 2021, 2023
- **ACL ARR** 2021-2023
- **NeurIPS** 2023

Workshops

- Workshop on Multimodal Content Moderation (MMCM) at CVPR 2023
- NLP4PI at ACL 2021

Volunteer: EACL 2021, EMNLP 2020, ACL2020, 2018 Singapore Symposium on NLP