Liwei Jiang

Paul G. Allen School of Computer Science & Engineering
University of Washington
Box 352350
Seattle, WA 98195

+1 207-313-6030
lwjiang@cs.washington.edu
https://liweijiang.me
@liweijianglw

RESEARCH FOCUS

My research interests center around **the co-evolution of AI and humanity**: developing AI systems with deep concerns for human *traits*, *values*, and *needs*, while gleaning novel insights into *humans*, *society*, and *humanity* through advancing AI.

EDUCATION

University of Washington, Seattle, WA

09/2019–current Ph.D. in Computer Science & Engineering, GPA 3.90
Advisor: Yejin Choi

Colby College, Waterville, ME *(top 1%, Dean's List 15–19)*

09/2015–01/2019 B.A. in Computer Science, *summa cum laude*, GPA 4.08
B.A. in Mathematics, *summa cum laude*, GPA 4.13

PROFESSIONAL EXPERIENCES

Allen Institute for Artificial Intelligence (Ai2)

06/2020–current Research Intern at the Mosaic Team, *with Yejin Choi*

University of Washington, Computer Science & Engineering

04/2020–current Research Assistant, *with Yejin Choi*
AI safety, computational morality, and alignment of human values and cultures

09/2019–08/2020 Research Assistant, *with James Fogarty*
Effective and easy-to-use self-tracking tools for migraine patients [P.12]

Stanford University, Computer Science

06/2017–09/2019 Research Intern, *with James Landay*
Educational interactive conversational systems, including *QuizBot* [P.31], *EnglishBot* [P.30], and *BookBuddy* [W.1]

Colby College, Computer Science

09/2017–01/2019 Undergraduate Research Assistant, *with Bruce Maxwell*
Honors thesis project on *Assistive Robot Guide for Visually Impaired Users*

05/2016–08/2016 Undergraduate Research Assistant, *with Ying Li*
Navigation with mobile phones without connectivity (e.g., GPS or Wi-Fi)

PUBLICATIONS

*, †denote equal contribution

Manuscripts and Pre-prints

P.1 **Liwei Jiang**, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jenny Liang, Sydney Levine, Jesse Dodge, Keisuke Sakaguchi, Maxwell Forbes, Taylor Sorensen, Jon Borchardt, Jack Hessel, Saadia Gabriel, Yulia Tsvetkov, Oren Etzioni, Maarten Sap, Regina Rini, and Yejin Choi. Can machines learn morality? The Delphi Experiment.
Under Review at Nature Machine Intelligence

P.2 **Liwei Jiang**, Kavel Rao*, Seungju Han*, Allyson Ettinger, Faeze Brahman, Sachin Kumar, Niloofar Mireshghallah, Ximing Lu1, Maarten Sap, Nouha Dziri, and Yejin Choi.
WildTeaming at Scale: From In-the-Wild Jailbreak Tactics to (Adversarially) Safer Models.
In submission

P.3 Seungju Han*, Kavel Rao*, **Liwei Jiang**†, Allyson Ettinger†, Yuchen Lin, Nathan Lambert, Nouha Dziri, and Yejin Choi. WildGuard: Open One-stop Moderation Tools for Safety Risks, Jailbreaks, and Refusals of LLMs.
In submission

P.4 Yu Ying Chiu, **Liwei Jiang**, Maria Antoniak, Chan Young Park, Shuyue Stella Li, Mehar Bhatia, Sahithya Ravi, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. CulturalTeaming: AI-Assisted Interactive Red-Teaming for Challenging LLMs' (Lack of) Multicultural Knowledge.
In submission

P.5 Wenting Zhao, Tanya Goyal, Yu Ying Chiu, **Liwei Jiang**, Benjamin Newman, Abhilasha Ravichander, Khyathi Chandu, Ronan Le Bras, Claire Cardie, Yuntian Deng, and Yejin Choi.
WildHallucination: Evaluating Long-form Factuality in LLMs with Real-World Entity Queries.
In submission

Peer-reviewed Conference and Journal Publications

2024 P.6 Jimin Mun, **Liwei Jiang**, Jenny Liang, Inyoung Cheong, Nicole DeCario, Yejin Choi, Tadayoshi Kohno, and Maarten Sap. Particip-AI: A Democratic Surveying Framework for Anticipating Future AI Use Cases, Harms and Benefits.
AIES 2024

P.7 Huihan Li, **Liwei Jiang**, Jena D. Huang, Hyunwoo Kim, Sebastin Santy, Taylor Sorensen, Bill Yuchen Lin, Nouha Dziri, Xiang Ren, and Yejin Choi. Culture-Gen: Revealing Global Cultural Perception in Language Models through Natural Language Prompting.
COLM 2024

P.8 Jaehun Jung, Ximing Lu, **Liwei Jiang**, Faeze Brahman, Peter West, Pang Wei Koh, and Yejin Choi. Information-Theoretic Distillation for Reference-less Summarization.
COLM 2024

P.9 Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, **Liwei Jiang**, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position Paper: A Roadmap to Pluralistic Alignment.
ICML 2024

P.10 Jaehun Jung, Peter West, **Liwei Jiang**, Faeze Brahman, Ximing Lu, Jillian Fisher, Taylor Sorensen, and Yejin Choi. Impossible Distillation: from Low-Quality Model to High-Quality Dataset & Model for Summarization and Paraphrasing.
NAACL 2024

P.11 Jillian Fisher, Ximing Lu, Jaehun Jung, **Liwei Jiang**, Zaid Harchaoui, and Yejin Choi. JAMDEC: Unsupervised Authorship Obfuscation using Constrained Decoding over Small Language Models.
NAACL 2024

P.12 Yasaman S Sefidgar, Carla L Castillo, Shaan Chopra, **Liwei Jiang**, Tae Jones, Anant Mittal, Hyeyoung Ryu, Jessica Schroeder, Allison Cole, Natalia Murinova, Sean A Munson, and James Fogarty. MigraineTracker: Examining Patient Experiences with Goal-Directed Self-Tracking for a Chronic Health Condition.
CHI 2024 🏆 **Outstanding Paper Award**

P.13 Linlu Qiu, **Liwei Jiang**, Ximing Lu, Melanie Sclar, Valentina Pyatkin, Chandra Bhagavatula, Bailin Wang, Yoon Kim, Yejin Choi, Nouha Dziri, and Xiang Ren. Phenomenal Yet Puzzling: Testing Inductive Reasoning Capabilities of Language Models with Hypothesis Refinement.
ICLR 2024 (Oral)

P.14 Peter West*, Ximing Lu*, Nouha Dziri*, Faeze Brahman*, Linjie Li*, Jena D. Hwang, **Liwei Jiang**, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, Benjamin Newman, Pang Wei Koh, Allyson Ettinger, and Yejin Choi. The Generative AI Paradox: What It Can Create, It *May* Not Understand.
ICLR 2024

P.15 Taylor Sorensen, **Liwei Jiang**, Jena Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, Chandra Bhagavatula, Maarten Sap, John Tasioulas, and Yejin Choi. Value Kaleidoscope: Engaging AI with Pluralistic Human Values, Rights, and Duties.
AAAI 2024

2023 P.16 **Liwei Jiang***, Kavel Rao*, Valentina Pyatkin, Yuling Gu, Niket Tandon, Nouha Dziri, Faeze Brahman, and Yejin Choi. What Makes it Ok to Set a Fire? Iterative Self-distillation of Contexts and Rationales for Disambiguating Defeasible Social and Moral Situations.
Findings of EMNLP 2023

P.17 Peter West, Ronan Le Bras, Taylor Sorensen, Bill Yuchen Lin, **Liwei Jiang**, Ximing Lu, Khyathi Chandu, Jack Hessel, Ashutosh Baheti, Chandra Bhagavatula, and Yejin Choi. NovaCOMET: Open Commonsense Foundation Models with Symbolic Knowledge Distillation.
Findings of EMNLP 2023

P.18 Seungju Han, Junhyeok Kim, Jack Hessel, **Liwei Jiang**, Jiwan Chung, Yejin Son, Yejin Choi, and Youngjae Yu. Reading Books is Great, But Not if You Are Driving! Visually Grounded Reasoning about Defeasible Commonsense Norms.
EMNLP 2023

P.19 Hyunwoo Kim, Jack Hessel, **Liwei Jiang**, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. SODA: Million-scale Dialogue Distillation with Social Commonsense Contextualization.
EMNLP 2023 🏆 **Outstanding Paper Award**

P.20 Nouha Dziri*, Ximing Lu*, Melanie Sclar*, **Liwei Jiang**†, Xiang Lorraine Li†, Bill Yuchen Lin†, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and Fate: Limits of Transformers on Compositionality.
NeurIPS 2023 (Spotlight)

P.21 Ximing Lu, Faeze Brahman, Peter West, Jaehun Jang, Khyathi Chandu, Abhilasha Ravichander, Lianhui Qin, Prithviraj Ammanabrolu, **Liwei Jiang**, Sahana Ramnath, Nouha Dziri, Jillian Fisher, Bill Yuchen Lin, Skyler Hallinan, Xiang Ren, Sean Welleck, and Yejin Choi. Inference-Time Policy Adapters (IPA): Tailoring Extreme-Scale LMs without Fine-tuning.
EMNLP 2023

P.22 Yiming Zhang, Sravani Nanduri, **Liwei Jiang**, Tongshuang Wu, and Maarten Sap. BiasX: "Thinking Slow" in Toxic Content Moderation with Explanations of Implied Social Biases.
EMNLP 2023

P.23 Valentina Pyatkin, Jena D. Hwang, Vivek Srikumar, Ximing Lu, **Liwei Jiang**, Yejin Choi, and Chandra Bhagavatula. ClarifyDelphi: Reinforced Clarification Questions with Defeasibility Rewards for Social and Moral Situations.
ACL 2023

2022 P.24 Ximing Lu, Sean Welleck, **Liwei Jiang**, Jack Hessel, Lianhui Qin, Peter West, Prithviraj Ammanabrolu, and Yejin Choi. Quark: Controllable Text Generation with Reinforced Unlearning.
NeurIPS 2022

P.25 Hyunwoo Kim*, Youngjae Yu*, **Liwei Jiang**, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. ProsocialDialog: A Prosocial Backbone for Conversational Agents.
EMNLP 2022

P.26 Prithviraj Ammanabrolu, **Liwei Jiang**, Maarten Sap, Hannaneh Hajishirzi, and Yejin Choi. Aligning to Social Norms and Values in Interactive Narratives.
NAACL 2022

P.27 Ximing Lu, Sean Welleck*, Peter West*, **Liwei Jiang**†, Jungo Kasai†, Daniel Khashabi†, Ronan Le Bras†, Lianhui Qin†, Youngjae Yu†, Rowan Zellers†, Noah A. Smith, and Yejin Choi. NeuroLogic A*esque Decoding: Constrained Text Generation with Lookahead Heuristics.
NAACL 2022 🏆 **Best Paper Award**

P.28 Peter West, Chandra Bhagavatula*, Jack Hessel*, Jena D. Hwang*, **Liwei Jiang***, Ronan Le Bras*, Ximing Lu*, Sean Welleck*, and Yejin Choi. Symbolic Knowledge Distillation: from General Language Models to Commonsense Models.
NAACL 2022

2021 P.29 **Liwei Jiang**, Antoine Bosselut, Chandra Bhagavatula, and Yejin Choi. "I'm Not Mad": Commonsense Implications of Negation and Contradiction.
NAACL 2021

P.30 **Liwei Jiang***, Sherry Ruan*, Qianyao Xu*, Zhiyuan Liu, Glenn M. Davis, Emma Brunskill, and James A. Landay. EnglishBot: An AI-Powered Conversational System for Second Language Learning.
IUI 2021

2019 P.31 Sherry Ruan, **Liwei Jiang**, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L. Murnane, Emma Brunskill, and James A. Landay. QuizBot: A Dialogue-based Adaptive Learning System for Factual Knowledge.
CHI 2019

Posters, Extended Abstracts, Workshop Papers and Technical Reports

2019 W.1 Sherry Ruan, Angelica Willis, Qianyao Xu, Glenn M. Davis, **Liwei Jiang**, Emma Brunskill, and James A. Landay. BookBuddy: Turning Digital Materials Into Interactive Foreign Language Lessons Through a Voice Chatbot.
L@S WIP 2019

HONORS AND AWARDS

2024 Outstanding Paper Award
CHI 2024

2023 Outstanding Paper Award
EMNLP 2023

2022 Best Paper Award
NAACL 2022

2019–2020 Anne Dinning - Michael Wolf Endowed Regental Fellowship
University of Washington, Paul G. Allen School First-Year Ph.D. Fellowship

2018 Member of the Phi Beta Kappa Society
Colby College, elected as a member of Phi Beta Kappa with junior standing

2016, 2017, 2018 Julius Seelye Bixler Scholar
Colby College, top-ranking students as determined by the academic record, three-time recipient

2018 Honorable Mention of Interdisciplinary Contest in Modeling (ICM)
20th annual Interdisciplinary Contest in Modeling (ICM)

2017 Phi Beta Kappa Undergraduate Scholastic Achievement Award
Colby College, top two students in the sophomore and junior classes

2016 Phi Beta Kappa Summer Research Scholar
Colby College, summer research stipend

INVITED TALKS

2024 **WildTeaming at Scale: From In-the-Wild Jailbreaks to (Adversarially) Safer Language Models** Allen Institute for Artificial Intelligence (AI2), All-AI2 Meeting (2024.7)

2023 **On the Outcomes of Scientific Disagreements on Machine Morality**
Co-speaker w/ Zeerak Talat, The Big Picture Workshop @ EMNLP, Singapore (2023.12)
Can we teach machines human ethics and values?

Co-speaker w/ Valentina Pyatkin and Taylor Sorensen, The Downtown School, Seattle (2023.9)
Toward Interpretable and Interactive Socially & Ethically Informed AI
Invited speaker, graduate seminar "AI Perspectives: Symbolic Reasoning to Deep Learning" at Computer Science Department at Northwestern University (2023.6)
Toward Interpretable and Interactive Socially & Ethically Informed AI
Speaker, Darpa ITM Kickoff PI Meeting (2023.5)
Toward Interpretable and Interactive Socially & Ethically Informed AI
Guest lecturer, LAW E 553 Technology Law And Public Policy Seminar at UW (2023.3)
Toward Interpretable, Interactive, Informative Machine Moral Reasoning
Discussant, AI2 Mosaic Morality & AI Series, AI2 (2023.2)

2022 **Toward Socially Aware & Ethically Informed AI**
UW NLP Retreat (2022.9)
Toward Socially Aware & Ethically Informed AI
Co-guest lecturer w/ Saadia Gabriel, The Downtown School, Seattle (2022.9)
Toward Ethically Informed & Socially Aware AI
Guest lecturer, HONORS 222 B at UW (2022.5)

2021 **Delphi: Toward Machine Ethics and Norms**
Allen Institute for Artificial Intelligence (AI2), All-AI2 Meeting (2021.10)

TEACHING EXPERIENCES

Teaching Assistant

01/2024–03/2024 **CSE447/517 Natural Language Processing**, UW
Head TA for the natural language processing class with 230+ undergraduate and graduate students
Co-design the class module, including teaching materials and homework

01/2023–03/2023 **CSE599 D1 Exploration on Language, Knowledge, and Reasoning**, UW
TA for a graduate seminar with 30+ students

09/2016–01/2019 **CS151 Introduction to Computational Thinking**, Colby College
CS231 Data Structure & Algorithm, Colby College
CS251 Data Analysis & Visualization, Colby College
Graded programming projects and homework, held TA office hours and tutored sessions weekly

09/2018–01/2019 **MA311 Ordinary Differential Equation**, Colby College
Held TA office hours and graded problem sets weekly for 30 students

Guest Lecturer

08/2024 **How can we build AI with deep concerns for human traits, values, and needs?**
In CSE 163: Intermediate Data Programming, University of Washington

09/2023 **Can we teach machines human ethics and values?**
In Ethics and Citizenship, w/ Valentina Pyatkin and Taylor Sorensen, The Downtown School, Seattle

03/2023 **Toward Interpretable and Interactive Socially & Ethically Informed AI**
In LAW E 553: Technology Law And Public Policy Seminar, University of Washington

09/2022 **Toward Socially Aware & Ethically Informed AI**
In Ethics and Citizenship, w/ Saadia Gabriel, The Downtown School, Seattle

05/2022 **Toward Ethically Informed & Socially Aware AI**
In HONORS 222 B: Artificial Intelligence Meets Society, University of Washington

MENTORING EXPERIENCES

Undergraduate & Master Students

01/2022–present **Kavel Rao** (Undergraduate student at UW CSE)
Explainable defeasible moral reasoning [P.16].
Open-source AI Safety Tool & In-the-wild safety redteaming [P.3, P.2].
🏆 Single Awardee of the Best Senior Thesis Award at UW (2024)

03/2023–present **Kelly Chiu** (Master student at UW Linguistics)
An AI-assisted interface for challenging multi-cultural quiz collection [P.4].

03/2023–07/2023 **Airei Fukuzawa** (Undergraduate student at UW CSE)
Enhancing LLMs with multi-cultural understanding and social norms.

09/2021–02/2023 **Sravani Nanduri** (Undergraduate student at UW CSE)
Co-supervised with Maarten Sap & Tongshuang (Sherry) Wu
Online hate speech moderation with explanations [P.22].

12/2021–03/2022 **Nuria Alina Chandra** (Undergraduate student at UW CSE)

Junior Graduate Students

05/2024–present **Jing-Jing Li** (PhD student at Berkeley)
Interpretable harm and benefit analysis of user queries to language models.

06/2023–02/2024 **Jimin Mun** (PhD student at CMU)
A democratic surveying framework for future AI harms and benefits [P.6].

06/2023–03/2024 **Huihan Li** (PhD student at USC)
Multicultural symbol generation and evaluation [P.7].

05/2023–09/2023 **Linlu Qiu** (PhD student at MIT)
Inductive reasoning capabilities of language models [P.13].

09/2022–08/2023 **Taylor Sorensen** (PhD student at UW CSE)
Engaging machines with pluralistic human values, rights, and duties [P.15].

01/2022–05/2023 **Jillian Fisher** (PhD student at UW Statistics)
Model revision and authorship obfuscation [P.11].

PROFESSIONAL SERVICE

Organizing Committees

2024 **Socially Responsible Language Modelling Research** (SoLaR Workshop, NeurIPS 2024)

2023 **AI Meets Moral Philosophy and Moral Psychology: An Interdisciplinary Dialogue about Computational Ethics** (MP2 Workshop, NeurIPS 2023)

Paper Reviewing

Conf. **EMNLP** 2022, **ACL** 2021, **AAAI** 2023, **NeurIPS** 2024, **NeurIPS D&B** 2024

Journal **Language Resources and Evaluation (Springer Nature)** 2024, **Journal of Experimental & Theoretical Artificial Intelligence** 2024

Community Service

2024 **Liaison**, UW Allen School Faculty Recruiting
Keep students informed about faculty recruiting and coordinate with student hosts to carry out responsibilities.

2023 **Area Chair, Reviewer**, UW Allen School PhD Admissions
Coordinate between PhD students/postdocs reading for PhD admissions and advising staff/admissions committee.
Liaison, UW Allen School Faculty Recruiting
Keep students informed about faculty recruiting and coordinate with student hosts to carry out responsibilities.
Co-chairs, UW Allen School Prospective Student Committee
Organize visit days for prospective PhD students at UW Allen School.

2022 **Volunteer**, UW NLP Retreat
Area Chair, Reviewer, UW Allen School PhD Admissions
Coordinate between PhD students/postdocs reading for PhD admissions and advising staff/admissions committee.
Liaison, UW Allen School Faculty Recruiting
Keep students informed about faculty recruiting and coordinate with student hosts to carry out responsibilities.
Student Representative, UW Allen School Diversity and Inclusion Committee
Co-chairs, UW Allen School Prospective Student Committee
Organize visit days for prospective PhD students at UW Allen School.
Volunteer Coordinator, NAACL 2022

2021 **Mentor**, UW Allen School Pre-Application Mentorship Service (PAMS)
A program supporting potential CS PhD applicants, with 80% from underrepresented communities.
Liaison, Reviewer, UW Allen School PhD Admissions

2020 **Co-organizer**, UW Allen School Pre-Application Mentoring Service (PAMS)

SELECTED MEDIA COVERAGE

2024 WildTeaming: An Automatic Red-Team Framework to Compose Human-like Adversarial Attacks Using Diverse Jailbreak Tactics Devised by Creative and Self-Motivated Users in-the-Wild
MarkTechPost, 7/2024

2023 How Moral Can A.I. Really Be?
The New Yorker, 11/2023
How Robots Can Learn to Follow a Moral Code
Nature Outlook, 10/2023

2022 Can Computers Learn Common Sense?
The New Yorker, 04/2022

2021 Can a Machine Learn Morality?
The New York Times, 12/2021

2021 This Program Can Give AI a Sense of Ethics–Sometimes
Wired, 12/2021

2021 Machines Learn Good From Commonsense Norm Bank
IEEE Spectrum, 11/2021

Updated August 2024