

RESEARCH INTEREST

Understand & steer the behavior of increasingly capable AI systems—encompassing capability development and scalable methods for alignment, oversight, and risk mitigation.

CONTACT

✉ <https://xudongolivershen.github.io>
 ✉ <https://www.linkedin.com/in/oliverxudongshen>
 ✉ xudong.shen@u.nus.edu
 ☎ (+65) 88296866 or (+86) 18658143615

EDUCATION

Ph.D. in Computer Science, *National University of Singapore*, Singapore 2019-2024
 B.A. in Naval. Arch. & Ocean Eng., *Zhejiang University*, China 2015-2019

PUBLICATIONS

On AI fairness:

1. **Xudong Shen**, Chao Du, Tianyu Pang, Min Lin, Yongkang Wong, Mohan Kankanhalli, “Fine-tuning Text-to-Image Diffusion Models for Fairness”, In *ICLR* (2024), (Oral, top 1.2%).
2. **Xudong Shen**, Yongkang Wong, Mohan Kankanhalli, “Fair Representation: Guaranteeing Approximate Multiple Group Fairness for Unknown Tasks”. In *IEEE Trans. PAMI* (2023).
3. **Xudong Shen**, Tianhui Tan, Tuan Q. Phan, Jussi Keppo, “Gender Animus Can Still Exist Under Favorable Disparate Impact: a Cautionary Tale from Online P2P Lending”. In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)* (2023).

On AI governance:

4. **Xudong Shen**, Hannah Brown, Jiashu Tao, Martin Strobel, Yao Tong, Akshay Narayan, Harold Soh, Finale Doshi-Velez, “Directions of Technical Innovation for Regulatable AI Systems”, In *Communications of the ACM*, (2024).
5. Ayse Gizem Yasar, Andrew Chong[†], Evan Dong[†], Thomas Krendl Gilbert[†], Sarah Hladikova[†], Roland Maio[†], Carlos Mougán[†], **Xudong Shen**[†], Shubham Singh[†], Ana-Andreea Stoica[†], Savannah Thais[†], “Integration of Generative AI in the Digital Markets Act: Contestability and Fairness from a Cross-Disciplinary Perspective”, (2024), LSE working papers series. ([†] = equal contribution)

On AI safety and benchmarking:

6. Paul Röttger, ..., **Xudong Shen**, ... (22 authors), “MSTS: A Multimodal Safety Test Suite for Vision-Language Models”, In *ArXiv* (2025).
7. Ian McKenzie, ..., **Xudong Shen**, ... (26 authors), “Inverse Scaling: When Bigger Isn't Better”, In *TMLR* (2023).
8. Kaustubh D Dhole, ..., **Xudong Shen**, ... (125 authors), “NL-Augmenter: A Framework for Task-Sensitive Natural Language Augmentation”. In *Northern European Journal of Language Technology (NEJLT)* (2023).
9. Aarohi Srivastava, ..., **Xudong Shen**, ... (450 authors), “Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models”. In *JMLR* (2023).
10. Yizhong Wang, ..., **Xudong Shen**, ... (40 authors), “Benchmarking Generalization via In-Context Instructions on 1,600+ Language Tasks”. In *EMNLP* (2022).

On AI capability development:

11. Ziwei Xu, **Xudong Shen**, Yongkang Wong, Mohan Kankanhalli, “Unsupervised Motion Representation Learning with Capsule Autoencoders”. In *NeurIPS* (2021).

POSITIONS

AI Research Lead, **Gata AI**, Singapore 2024-Current
 (a) Research on AI-driven data for AI training, decentralized inference & training
 (b) Helped raise \$60M seed round from top tier investors
Doctoral Researcher, **N-CRiPT Centre**, NUS, Singapore 2023–2024
 (c) Conduct research in ML and FinTech: (i) guarantee fairness in downstream tasks via pre-trained fair representations, (ii) structural modeling of million online peer-to-peer lending transaction data to estimate default risk, profitability, and discrimination.

- (d) Communicated scientific insights to both academic and industry audience in 10 papers (4 first-authored) and 5+ talks.
- (e) Assisted teaching the course *Computational Methods for Business Analytics* twice, totally 150+ hours to 200+ undergraduate students. Prepared weekly recap materials, videos, and Q&A sessions.
- (f) Actively contributed to open-community projects in NLP, including [BIG-Bench](#), [natural language instructions](#), [NL-Augmenter](#), and [LLM's Inverse Scaling](#).
- (g) Co-organized (i) the Responsible, Regutable AI working group at NUS and (ii) the [Algorithms, Law, and Policy working group](#) at [EAAMO bridges](#) (formerly [MD4SG](#)).

Research Intern, **Sea AI Lab**, Sea Limited, Singapore Nov. 2022–Sep. 2023

- (a) led collaboration with 5 researchers from 2 institutes (NUS & Sea AI Lab) on finetuning text-to-image diffusion models for fairness. Oral presentation at ICLR 2024.

Undergrad Researcher, **Second Institute of Oceanography**, China Jul. 2017–Jul. 2019

- (a) Worked on the full stack of satellite remote sensing image segmentation using Python & Matlab, incl. data preprocessing, labeling, neural net training and validating.

Undergrad Researcher, **Waseda University**, Japan Sep. 2018–Feb. 2019

- (a) Analyzed the effectiveness of Japan's policy toward the elimination of Persistent Organic Pollutants (POPs), and proposed policy recommendations to China.

Undergrad Researcher, **York University**, Canada May–Aug. 2018

- (a) Worked on fluid dynamics simulation of a hydrodynamically focused printing process for printed electronics, using COMSOL and FLUENT, in collaboration with 5 other researchers.

SKILLS

AI Alignment:

Familiar with (i) the technical methods for ensuring the social alignment of AI (LLMs, text-to-image diffusion models,...) and (ii) AI governance.

ML and AI:

Experienced with (i) vision, NLP, generative, and multimodal models; (ii) image, sensor, sequential, language data, and (3) FinTech data.

Statistical Modeling:

Big data analysis and structural modelling using the R and Python stack. Experienced with survival modelling, financial default analysis, and social network data analysis.

TALKS

ICLR conference 2024, Vienna, Austria 2024
 SEA AI Lab, Singapore 2023, 2024
 N-CRiPT Seminar, *National University of Singapore*, Singapore 2022, 2023, 2024
 FAccT conference 2023, Chicago, USA 2023
 International Conference on Smart Finance 2022, Online 2022
 Workshop on Enormous Language Models, *ICLR conference 2021*, Online 2021

SERVICE

Organizer: [MD4SG Algorithms, Law, and Policy working group](#) AY 2023/2024.
Area Chair: NeurIPS 2023 Regulatable ML Workshop.
Reviewer: EMNLP 2022, ICLR 2023, WACV 2023, ACL 2023.

FELLOWSHIPS

DAAD Alnet Fellow, *DAAD*, Germany 2024
 Master Kong Dream Scholarship Program, *Waseda University*, Japan Sep. 2018-Feb. 2019
 Globalink Research Internship, *York University*, Canada May.-Aug. 2018
 Erasmus+ Student Mobility, *Università di Trento*, Italy Feb.-Jun. 2017