

CONTACT	<p>✉ xudong.shen@u.nus.edu ↵ Links: LinkedIn, Homepage, Google Scholar</p>	
INTRO	I am passionate about scaling Reinforcement Learning for multi-modal, long-horizon, complex real-world tasks . My earlier work focused on AI fairness, robustness, safety, and governance, where I took an evaluation-driven approach: stress-testing systems and developing model improvements.	
EDUCATION	<p>Ph.D. in Artificial Intelligence, National University of Singapore B.A. in Naval Architecture & Ocean Engineering, Zhejiang University</p>	2019–2024 2015–2019
EXPERIENCE	<p><i>Co-founder, Gata</i>, Singapore</p> <ul style="list-style-type: none"> – Led the company 0 to 1; raised seed funding. – Launched GPT-to-Earn data product: 15K+ users; 3.5M+ conversations collected. – Developing decentralized inference: optimizing throughput & end-to-end latency. <p><i>Research Intern, Sea AI Lab</i>, Sea Limited (NYSE: SE), Singapore</p> <ul style="list-style-type: none"> – Developed a method to optimize diffusion models for any differentiable objective (e.g., aesthetics, robustness); where score/flow matching and RL fail. – ICLR 2024 Oral; patents in US and China. <p><i>PhD, National University of Singapore</i>, Singapore</p> <ul style="list-style-type: none"> – Developed interpretable & robust representation learning methods. – Large-scale LLM/VLM eval on capabilities, safety, scaling behavior, & in-context learning. – Time-to-event modeling on million-scale lending data: modeled repayment as a survival process to forecast default risk and profitability. 	2024–2025 2022–2024 2019–2024
PUBLICATIONS	<p>Controllable Optimization for Generative Models</p> <p>1. <u>Xudong Shen</u>, Chao Du, Tianyu Pang, Min Lin, Yongkang Wong, Mohan Kankanhalli, “Finetuning Text-to-Image Diffusion Models for Fairness”, In ICLR (2024), (Oral, top 1.2%). <i>TLDR: We developed a method to optimize diffusion models for any differentiable objective defined on the generated data, where score/noise prediction and RL fail. We applied it to control output diversity in text-to-image generation.</i></p> <p>Foundation Model Evaluations & Training</p> <p>2. Ian McKenzie, ..., <u>Xudong Shen</u>, ... (26 authors), “Inverse Scaling: When Bigger Isn’t Better”, In TMLR (2023). <i>TLDR: Shows when larger models consistently perform worse; analyzes failure modes.</i></p> <p>3. Aarohi Srivastava, ..., <u>Xudong Shen</u>, ... (450 authors), “Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models”. In TMLR (2023). <i>TLDR: Large-scale eval that reveals where LLM capabilities scale well & where they don’t.</i></p> <p>4. Kaustubh D Dhole, ..., <u>Xudong Shen</u>, ... (125 authors), “NL-Augmenter: A Framework for Task-Sensitive Natural Language Augmentation”. In NEJLT (2023). <i>TLDR: Stress-tested LLM robustness using 100+ natural-language augmentations.</i></p> <p>5. Yizhong Wang, ..., <u>Xudong Shen</u>, ... (40 authors), “Benchmarking Generalization via In-Context Instructions on 1,600+ Language Tasks”. In EMNLP (2022). <i>TLDR: Instruction-tuning on 1.6K tasks boosts zero-shot unseen-task performance.</i></p> <p>Safety & Bias Test Suites for LLM/VLMs</p> <p>6. Paul Röttger, ..., <u>Xudong Shen</u>, ... (22 authors), “MSTS: A Multimodal Safety Test Suite for Vision-Language Models”, In ArXv (2025). <i>TLDR: Multimodal safety test: image+text prompts trigger more safety failures than text-only.</i></p>	

7. Margaret Mitchell, ..., **Xudong Shen**, ... (55 authors), "SHADES: Towards a multilingual assessment of stereotypes in large language models", In **NAACL** (2025).
TLDR: *Probes multilingual stereotypes and its cross-lingual transfer in LLMs.*

Robust & Interpretable Representations

8. **Xudong Shen**, Yongkang Wong, Mohan Kankanhalli, "Fair Representation: Guaranteeing Approximate Multiple Group Fairness for Unknown Tasks". In **IEEE Trans. PAMI** (2023).
TLDR: *Learns representation with robustness guarantees that transfer to unseen tasks.*
9. Ziwei Xu, **Xudong Shen**, Yongkang Wong, Mohan Kankanhalli, "Unsupervised Motion Representation Learning with Capsule Autoencoders". In **NeurIPS** (2021).
TLDR: *Learns representations with built-in interpretability via capsule networks.*

Predictive Modeling for Real-World Decision Making

10. **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 **FAccT** (2023).
TLDR: *Time-to-event modeling to predict default & profitability on million-scale lending data.*

Regulatable AI: Policy & Technical Mechanisms

11. **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).
TLDR: *maps technical mechanisms that make AI easier to regulate in practice.*
12. Ayse Gizem Yasar, Andrew Chong[†], Evan Dong[†], Thomas Krendl Gilbert[†], Sarah Hladikova[†], Roland Maio[†], Carlos Mougan[†], **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**.
TLDR: *analyzes how GenAI interacts with platform regulation.*

OTHER EXPERIENCE	<i>Undergrad Researcher, Second Institute of Oceanography, China</i>	Jul. 2017–Jul. 2019
	– Trained CNNs for satellite remote sensing image segmentation.	
	<i>Undergrad Researcher, Waseda University, Japan</i>	Sep. 2018–Feb. 2019
	– Analyzed the effectiveness of Japan's policy toward the elimination of Persistent Organic Pollutants (POPs), and proposed policy recommendations to China.	
	<i>Undergrad Researcher, York University, Canada</i>	May–Aug. 2018
	– Worked on fluid dynamics simulation of a hydrodynamically focused printing process for printed electronics, using COMSOL and FLUENT, in collaboration with 5 other researchers.	
FELLOWSHIPS	DAAD Alnet Fellow, <i>DAAD, Germany</i>	2024
	Master Kong Dream Scholarship Program, <i>Waseda University, Japan</i>	Sep. 2018–Feb. 2019
	Globalink Research Internship, <i>York University, Canada</i>	May.–Aug. 2018
	Erasmus+ Student Mobility, <i>Università di Trento, Italy</i>	Feb.–Jun. 2017