Yuhong Luo yuhongluo@umass.edu

Statement of Purpose

My research interest is in ensuring the reliability and trustworthiness of AI, with a specific focus on fairness and robustness of complex machine learning (ML) systems.

Before I began the M.S. program at UMass, I worked as an engineer at Airbnb on marketing content recommendations. My experience addressing positional bias in recommender systems deepened my understanding of modern ML's limitations. I learned that ML can exhibit harmful algorithmic biases, resulting in unfairness in critical domains (Obermeyer et al., 2019), fostering the spread of fake news within echo chambers in social networks (Cinelli et al., 2021), and making the system susceptible to adversarial attacks (Mehrabi et al., 2021). Since modern ML systems have become increasingly complex, many underlying issues continue to persist despite industry efforts to address these biases (Bender et al., 2021). My knowledge of these biases fuels my interest in studying the **reliability and trustworthiness of modern ML systems**.

Driven to explore emerging technologies for enhancing the fairness and robustness of recommender systems in e-commerce and social networks, I became particularly excited about graph representation learning for its role in anomaly detection (Ma et al., 2023). While I still worked at Airbnb, I initiated research with Prof. Pan Li of Georgia Tech, focusing on enhancing the expressiveness and scalability of temporal graph learning.

I found that a provably expressive approach to temporal graph learning by Wang et al. (2021) suffers from its slow process in training and inference, severely limiting its scalability. To mitigate this issue, I conducted fine-grained profiling of the learning procedures for multiple approaches to temporal graph learning (Wang et al., 2021; Kumar et al., 2019; Rossi et al., 2020), and identified a promising lead: many recent methods need to trace back in time and sample historical neighbors for learning a node representation. Unfortunately, these operations are slow. I brainstormed, explored, and eventually arrived at a dynamic-programming-inspired GPU-executable algorithm, which significantly outperforms baselines in terms of both *prediction performance and scalability*. We published the result (**Luo** & Li, 2022) at the *Learning on Graphs (LoG)* conference, and received a *best paper award*. From this experience, I have developed a strong passion for studying open-ended problems and improving complex ML systems, and am committed to pursuing a research career.

Eager to explore more research opportunities in the area of reliable and trustworthy AI, I joined the M.S. CS program at UMass. I was excited to work with Prof. Hui Guan and her team, focusing on improving the accuracy of personalized federated learning which preserves client privacy while achieving personalized predictions. I approached this problem as the adaptation of a foundation model to achieve personalized client models, but encountered challenges in providing equitable performance among clients with data distributions under statistical heterogeneity (Li et al., 2020). This also led me to discover current challenges in ensuring the fairness and robustness of all downstream tasks of the foundation models (Bommasani et al., 2022), whose solutions can provide insights for ensuring fair client performance in personalized federated learning. I was drawn toward exploring potential improvements that could enhance the fairness and robustness of foundation models and their adaptation.

I noticed that the backbone of foundation models consists of learned data representations that are valuable for various downstream prediction tasks. However, providers who distribute the foundation models or data representations may have limited control over their use in downstream models (Zemel et al., 2013; Madras et al., 2018). This realization allowed me to value the importance of ensuring the fairness of learned representations in a way that can ensure fairness for all downstream predictive tasks that use the representations. To delve deeper into this subject, I embarked on a project on **fair representation learning** in collaboration with Prof. Philip S. Thomas at UMass and Dr. Austin Hoag from Berkeley Existential Risk Initiative.

I discovered a knowledge gap in current fair representation learning methods: many prior methods lack high-confidence guarantees that the learned representations will remain fair for future unseen data and unfamiliar downstream tasks. To address this gap, I developed a framework capable of

offering **high-confidence fairness guarantees in representation learning** and supported this claim with both *theoretical and empirical analysis*. This paper (**Luo** et al., 2023) is currently in submission. I plan to continue this line of work to further explore high-confidence guarantees for the fairness and robustness of representation learning under distributional shifts and the adaptation of foundational models.

With my growing confidence and affinity in exploring and bridging knowledge gaps to advance the frontier of knowledge, I am dedicated to enhancing the reliability and trustworthiness of AI through research. **For my next step**, I aim to pursue a PhD to enhance my research expertise and further refine my research focus. I look forward to joining research labs at Columbia and learning about AI trustworthiness from different perspectives. While I am still exploring various aspects of AI trustworthiness, there are a couple of directions that stand out to me as ones that could make for interesting projects.

Fairness Guarantees Under Distributional Shifts and Causality. During my research on providing high-confidence fairness guarantees for representation learning, I learned that classical statistical tools may be inadequate in guaranteeing fairness in the presence of certain distributional shifts, such as covariate shifts (Singh et al., 2021). Conducting a statistical test on a holdout test set is insufficient to ensure the validity of the results when experimental conditions change. Additionally, these distributional shifts significantly affect the robustness and generalizability of ML models (D'Amour et al., 2020; Wang et al., 2023). Several studies have suggested that these problems stem from the absence of causal formalisms (Pearl, 2019; Kaddour et al., 2022), and have proposed incorporating graphical causality to distinguish shift-invariant causal factors from spurious correlations, including tasks like causal disentanglement of representations (Yang et al., 2021), invariant feature learning (Sun et al., 2021; Lu et al., 2022), and causal domain adaptation (Singh et al., 2021). I am curious *under what circumstances (e.g., data distributions and availability) can the identifiability of the fairness constraints be guaranteed, how to construct fairness guarantees using causal formalisms, and how generalizable they are under distributional shifts.*

Robust Adaptation of Foundation Models. Today, many large-scale ML systems are transitioning to training foundation models on extensive datasets and then adapting these models through fine-tuning to various downstream tasks across diverse domains (e.g., GPT-4 (OpenAI, 2023)). Some recent work has shown that foundation models can improve robustness to distribution shifts due to their training on diverse data encompassing a wide distribution (Radford et al., 2021). However, their robustness is limited, particularly when fine-tuning the foundation model for few-shot learning because it can distort the pre-trained feature extractor (Xie et al., 2021; Kumar et al., 2022). Wortsman et al. (2022) showed that ensembling the weights of the zero-shot and fine-tuned models can lead to better accuracy while maintaining robustness. It is also suggested that there exists a model that leverages the foundation model's robustness and enhances accuracy through fine-tuning. Drawing inspiration from the lottery ticket hypothesis (Frankle & Carbin, 2019), I am interested in gaining fundamental understanding of robust adaptation by exploring *whether it is possible to identify sub-networks within the foundation model that contribute to robustness, how the subnetworks change with fine-tuning, and how these sub-networks may vary across downstream tasks.*

Columbia University's Computer Science PhD program stands out to me for several reasons. First, the faculty's expertise on trustworthy AI is exceptional. I am eager to work with **Richard Zemel** on fairness and robustness in modern AI systems; with **Carl Vondrick** on adversarial robustness for large-scale models (Mao et al., 2023a,b); and with **David Blei** and **Elias Bareinboim**, as their expertise in Bayesian methods and causal inference forms a crucial foundation for ensuring fairness and robustness in AI. Columbia's rich culture for interdisciplinary collaboration is another aspect that invigorates me. I am also interested in collaborating with experts in other fields, such as the Decision, Risk, and Operations (DRO) of the business school, statistics and data science, as I look forward to opportunities for analyzing AI trustworthiness in diverse domains.

In the future, I aim to apply the research expertise I develop during my PhD to address research questions concerning the fairness, robustness and trustworthiness of cutting-edge technologies. I also aspire to collaborate with industry researchers to bring these ideas into fruition.

References

Yuhong Luo and Pan Li. Neighborhood-aware scalable temporal network representation learning. In *Learning on Graphs*, 2022.

Yuhong Luo, Austin Hoag, and Philip S. Thomas. Learning fair representations with high-confidence guarantees. *arXiv preprint arXiv:2310.15358*, 2023.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–623, 2021.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2022.

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118, 2021.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *ICLR*, 2019.

Jean Kaddour, Aengus Lynch, Qi Liu, Matt J. Kusner, and Ricardo Silva. Causal machine learning: A survey and open problems. *arXiv preprint arXiv:2206.15475*, 2022.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *ICLR*, 2022.

Srijan Kumar, Xikun Zhang, and Jure Leskovec. Predicting dynamic embedding trajectory in temporal interaction networks. In *KDD*, 2019.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, and Bernhard Schölkopf. Invariant causal representation learning for out-of-distribution generalization. In *International Conference on Learning Representations*, 2022.

Xiaoxiao Ma, Jia Wu, Shan Xue, Jian Yang, Chuan Zhou, Quan Z. Sheng, Hui Xiong, and Leman Akoglu. A comprehensive survey on graph anomaly detection with deep learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12012–12038, 2023.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3384–3393. PMLR, 2018.

Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. In *ICLR*, 2023a.

Chengzhi Mao, Lingyu Zhang, Abhishek Joshi, Junfeng Yang, Hao Wang, and Carl Vondrick. Robust perception through equivariance. In *ICML*, 2023b.

Ninareh Mehrabi, Muhammad Naveed, Fred Morstatter, and Aram Galstyan. Exacerbating algorithmic bias through fairness attacks. In *AAAI*, 2021.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019. doi: 10.1126/science.aax2342.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Judea Pearl. The seven tools of causal inference, with reflections on machine learning. *Commun. ACM*, 62(3):54–60, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael Bronstein. Temporal graph networks for deep learning on dynamic graphs. In *ICML 2020 Workshop on GRL*, 2020.

Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021.

Xinwei Sun, Botong Wu, Xiangyu Zheng, Chang Liu, Wei Chen, Tao Qin, and Tie-Yan Liu. Recovering latent causal factor for generalization to distributional shifts. In *Advances in Neural Information Processing Systems*, 2021.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip S. Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 35(8):8052–8072, 2023.

Yanbang Wang, Yen-Yu Chang, Yunyu Liu, Jure Leskovec, and Pan Li. Inductive representation learning in temporal networks via causal anonymous walks. In *ICLR*, 2021.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Hanna Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. In *CVPR*, 2022.

Sang Michael Xie, Tengyu Ma, and Percy Liang. Composed fine-tuning: Freezing pre-trained denoising autoencoders for improved generalization. In *ICML*, 2021.

Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: Disentangled representation learning via neural structural causal models. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9588–9597, 2021.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *ICML*, 2013.