Introduction: I aspire to build an impactful research career focused on five foundational pillars: **Privacy**, **Fairness**, **Interpretability**, **Robustness**, **Explainability** for AI/ML Algorithms. Studying the interplay between these principles and working towards strengthening these grounding ideals for the large-scale deployment of ML systems worldwide is imperative for creating systems that are **accountable**, **accessible**, **equitable** and **inclusive**. My experiences as a Pre-Doctoral Researcher at Google Research India (GRI) along with my rigorous academic training at Indian Institute of Technology (IIT) Hyderabad (**CGPA: 9.79, Institute Rank: 2**) have prepared me for a research career that helps me tackle these key questions.

Motivation: Currently the black-box systems used for high-stakes decisions like criminal justice and even the now universally used LLMs suffer from unseen biases which have far-reaching and life-altering implications [1] There are strong regulatory headwinds [2] and technical leaders [3], including pioneers of the AI field, urging to develop models that are innately robust, interpretable, and safe, across and independent of their modalities. I strongly wish to equip myself with expertise in the above mentioned five bastions to pioneer the future of AI. I plan to create systems driven by these traits and enable more people to do the same. My experience at Google's influential Advertising Sciences Team instilled a commitment to coding excellence, motivating me to transform ideas into large-scale functional systems. I'm passionate about exploring new domains and working on high-impact projects beyond these spheres. My work at Google under the guidance of Dr. Aravindan Raghuveer, Dr. Karthikeyan Shanmugam & Dr. Rishi Saket in the past couple of years, graduation from IIT Hyderabad, with a Bachelor of Technology in Computer Science (Honors), where I primarily worked with Prof. Vineeth N. Balasubramanian, and my internships at Adobe and JIW have helped me gain essential skills and actualize my research interests, to embark on this journey, which I further describe in detail below.

Privacy & Fairness: At GRI, I worked on the Learning from Aggregate Data problem, focusing on Learning from Label Proportions (LLP), where only aggregate level training labels are available for groups of instances (bags), while one evaluates on instance-level test data. This setting arises in domains like advertising and medicine due to privacy considerations and is of great significance to Google's Advertising objectives. I led a project to develop a two-step efficient and scalable algorithmic framework utilizing *Belief Propagation* and *Embedding Refinement*, resulting in a submission to ICLR '24 receiving positive reviews and a highly appreciated **Oral** at Regulatable ML @ NeurIPS2023 [4] The availability of training labels only in the form of aggregates, (highly limited supervision) makes it difficult to obtain good test performance. One has to explicitly leverage the fact that similar covariates must have similar labels. But even the right representation is not apriori available to enforce this directly. To identify and tackle these challenges, I noticed the connection to coding theory, where one of the fundamental problems is to decode an unknown message string sent by the encoder using only parity checks over groups of bits from the message. I drew the parallel to the aggregate label and covariate similarity enforcing constraints. This unique perspective helped develop a fresh approach to an increasingly important problem in light of new and much required equality and privacy enhancing regulations. Our algorithm displays strong gains against several SOTA baselines (up to **15%**) for the LLP Binary Classification problem on various dataset types - tabular and image, for large bag sizes, even for a million samples. I took the lead on all phases of the research project cycle, including algorithmic realization of this intuitive solution, efficient implementation, extensive analysis and paper writing. I also gave talks disseminating the work, getting invaluable feedback, learning from diverse opinions. This helped me appreciate and understand the end-to-end research process. This project helped me understand the importance of collaboration, and how having diverse viewpoints can generate effective solutions. It also helped me grasp the interdisciplinary nature of research, and how to connect the dots across literature. Finally, the experience taught me the value of persistence and practicality in making an idea work.

While this project helped me formulate an empirical approach to the LLP problem, I also wanted to explore its theoretical implications through the formal lens of privacy. I formulated new problems in attainable levels of differential privacy through aggregation methods leading to an ongoing submission to ICML [5], thus broadening my horizons.

In keeping with my overall goal of developing fair and deployable models, while at GRI, I also worked on analyzing current fairness methods for their performance on covariate shifts. I observed the catastrophic consequences on a SOTA method's train set fairness-accuracy tradeoff when evaluated on a covariate shifted test set. I noticed the lack in solutions utilizing a few unlabeled test samples, and ideated a novel approach, comprising a composite weighted entropy based objective for prediction accuracy which is optimized along with a representation matching loss for fairness. Backed by theoretical reasoning, the method outperformed a number of state-of-the-art baselines in the pareto sense with respect to the fairness-accuracy tradeoff on several standard datasets and culminated in highly praised **AAAI '24** main paper and a **Spotlight** at Algorithmic Fairness through the Lens of Time @ NeurIPS 2023 [6] My analysis of advertising datasets revealed the existence of a yet unexplored, yet important problem, I coined *Asymmetric Covariate Shift*. The distribution of covariates of one group (say Small and Medium Business Advertisers) shifts significantly compared to the other groups when a dominant group (Large Advertisers) is over-represented. While this setting is extremely challenging for current baselines, our proposed method significantly outperforms them. Performing the empirical analysis to posit not only a new real-world setting but also a solution for it that has been appreciated by the community as "*realistic and a good test to evaluate models in future*", has imbibed in me the aptitude for not only identifying and closing gaps in literature but also opening up new directions of research.

Robustness, Interpretability & Explainability: Since my sophomore year at IIT Hyderabad, my devouring of online resources on ML, strong mathematical and CS foundations, and my curiosity to delve deeper drew me to the Department Head of AI; Prof. Vineeth's Lab 1055, The Machine Learning and Vision Lab at IIT Hyderabad. Most of my undergraduate research effort concentrated on Interpretability & Robustness. We hypothesized that the underlying hidden structure within the data naturally follows a linear pattern, making the latent space of a generative model better suited for establishing vicinal distributions through linear interpolations like MixUp instead of sampling images from the feature space. My empirical verification of this hypothesis resulted in inherently more robust, better calibrated models with more local-linear loss landscapes. This resulted in a **Best Paper Award** at AML-CV Workshop, CVPR '21 and 2 further workshops in ICLR '21 [7]

Exploration: Concurrently, I dabbled in a multitude of domains, from Quantum Algorithms with Prof. M.V. Panduranga Rao, to my Honors Thesis on Union-Closed Conjecture with Prof. Rogers Mathew, Privacy in Deep Learning Systems with Prof. Maria Francis and even Algorithmic Complexity for Games with Prof. Karteek Sreenivasaiah, presenting my work to a heterogeneous audience. I have constantly been motivated by a research career with experiences spanning Adobe Research, Bengaluru on improving document tagging for enhancing accessibility by incorporating human feedback, and JIW, Tokyo to develop aerial-imagery based semantic segmentation models. Learning the faculty to develop and present live demos to large leadership audiences and working with people from myriad backgrounds and nationalities helped me gain invaluable insights on working together towards a common goal, overcoming language, time zone and cultural barriers.

Along the way, involvement in teaching for courses like Foundations of ML, Complexity Theory, Data Structures, Artificial Intelligence at, Google, IIT Hyderabad and National Programme on Technology Enhanced Learning (NPTEL), have significantly fueled my desire to pursue a PhD and embark on a research path. I've always been a strong believer in giving back to the community, be it via tutoring the underprivileged or developing material for the disadvantaged, or simply acting as reviewer and volunteer for conferences. I am passionate about mentoring students, probing deep into problems and solving them end to end and hence a Ph.D would elevate my current foundational skills and provide a launchpad towards a successful research career.

Looking forward: I aim to innovate solutions to empower the AI systems of tomorrow to be safe, intelligible and fair. The potential to create profound impact via ML has never been higher, and so has the responsibility. *Can we develop ML algorithms which can explain their own explanations for their decisions? Can we infuse into our large-scale systems the resilience to malicious attacks? Can we ensure a user of the AI Tool that their data shall be privy to no other user or* *administrator, or even enable the model to unlearn it if they chose so? Can we ensure each user, irrespective of which corner of the globe they come from, are viewed from an unbiased lens by the methods we develop? Can we confidently distinguish an image/article by a Generative Model from a natural or human attributed one via watermarking, nurturing trustworthy ML?* A PhD provides me the platform to be guided by the best minds in the field, and having the academic freedom to stay onto difficult questions and study them objectively with an aim to significantly advance the field.

Columbia is a cradle of innovation, pushing the frontiers of research and real-world applications, exemplifying why it is the perfect place to call my home for the next few years. I'm highly interested in working with the illustrious group at Columbia, as they align well with my five-pillared goal of Trustworthy ML. I would love to work with **Prof. Richard Zemel**, **Prof. Elias Bareinboim**, **Prof. David Blei**, **Prof. Carl Vondrick** to name a few. I look forward to learning from and work with many such pioneers at Columbia's illustrious research groups.

Thus, I believe that a PhD at Columbia propels me to effectively further my research ambitions and enables me to give back to the community at large, be it via teaching, mentoring or even simply opening up new avenues of research that can become the spark for someone else like me to pursue a few years down the line.

References

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Machine bias". In: *Ethics of data and analytics.* Auerbach Publications, 2022, pp. 254–264.

[2] Joseph R Biden. "Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence". In: (2023).

[3] Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, et al. "Managing AI Risks in an Era of Rapid Progress". In: *arXiv preprint arXiv:2310.17688* (2023).

[4] Shreyas Havaldar, Navodita Sharma, Shubhi Sareen, Karthikeyan Shanmugam, and Aravindan Raghuveer. "Learning from Label Proportions: Bootstrapping Supervised Learners via Belief Propagation". In: *arXiv preprint arXiv:2310.08056* (2023).

[5] Anand Brahmbhatt, Rishi Saket, Shreyas Havaldar, Anshul Nasery, and Aravindan Raghuveer. "Label Differential Privacy via Aggregation". In: *arXiv preprint arXiv:2310.10092* (2023).

[6] Shreyas Havaldar, Jatin Chauhan, Karthikeyan Shanmugam, Jay Nandy, and Aravindan Raghuveer. "Improving Fairness-Accuracy tradeoff with few Test Samples under Covariate Shift". In: *arXiv preprint arXiv:2310.07535* (2023).

[7] Puneet Mangla, Vedant Singh, Shreyas Havaldar, and Vineeth Balasubramanian. "On the benefits of defining vicinal distributions in latent space". In: *Pattern Recognition Letters* 152 (2021), pp. 382–390.