ishapuri@college.harvard.edu ishapuri.github.io

MIT IDSS, Statement of Purpose: Isha Puri

Ever since I was first introduced to AI through Professor Fei-Fei Li's inaugural Stanford AI Lab Summer Program (SAILORS) in 2016, my fascination with the field has been rooted in its tremendous ability for societal impact. My subsequent experiences have solidified **my strong research interest in Machine Learning and Natural Language Processing - explored through the powerful lens of explainable AI and reasoning** I am fortunate to have conducted research in these areas at IBM Research AI and Harvard (resulting in publications at NeurIPS & IEEE EMBC, and honors such as the ACM Cutler-Bell Prize for Computing and the NCWIT Collegiate National Award, among others), and I am thrilled to expand my research into these areas and explore new directions in my PhD.

I had my first immersive experience in the world of AI research when I joined the lab of Professor David Cox - then a Harvard professor, now Director of the MIT-IBM Watson AI Lab. Professor Cox tasked me with simplifying and improving the process used to collect animal subject eye movement data, a complex problem due to the presence of unique features in rodent eye images, e.g., variability in eye parameters, abundance of surrounding hair, and their small size. To overcome these unique challenges, **I developed a highly accurate (over 98% accuracy) and practical method leveraging biomedical image segmentation based CNN architecture**, which yielded a state of the art eye tracking capability for neuroscience and vision research with animal subjects. I published this work at IEEE's Engineering in Medicine and Biology Conference [1]. I subsequently used the domain familiarity I gained with eye tracking to create DYSCERN [2], an application that uses an inbuilt computer webcam to diagnose children for dyslexia with an accuracy of 90.18%. Dyslexic students exhibit unique eye movement patterns while reading, and in order to adapt to the constraints of low-quality video footage, I developed a machine learning based approach to track pupil movements across time. This solution provides the first-ever freely available, highly accurate test for risk of dyslexia, accessible to anyone around the world without regard to financial status or physical location.

My experience with DYSCERN showed me firsthand the impacts AI could have on society and the need for the study of explainable AI. To explore this area further, I started an internship at IBM Research AI's Trustworthy AI team, where I had the opportunity to work with Drs. Kush Varshney and Amit Dhurandhar. AI explainability is crucial, and yet most such methods are "post-hoc", taking place after the training and deployment of models. In my research, **I led the creation of CoFrNets (Continued Fraction Nets), an inherently interpretable neural network architecture based on the structure of continued fractions** Continued fractions are infinitely nested fractions with tremendous capabilities to approximate real numbers and functions [3]. I explored new ideas, wrote and debugged the code and iterated on experiments for the new architecture from beginning to end. In our NeurIPS 2021 paper [4], I found that our novel architecture was stable, low-resource, and performed with higher accuracy on tabular, image, and text datasets than its interpretable competitors [5][6]. I showed that CoFrNets demonstrate both local and global interpretability and mathematically proved their universal approximation ability. CoFrNets was patented by IBM, and I open-sourced CoFrNets as part of IBM's AI Explainability 360 (AIX360) [7] - a widely used toolkit (with more than 1.2k stars on GitHub) that puts CoFrNets into the hands of thousands of researchers and developers. This research taught me the importance of persistence in the research process, and the joy of this research journey solidified my desire to pursue a PhD. I continued my research in novel neural architectures through an internship at the Quantum AI Research team at IBM Research Zurich, implementing Quantum GAN (Generative Adversarial Networks) algorithms in QisKit - IBM's quantum open-sourced library.

My experience at IBM Research AI motivated me to continue my journey into explainable AI, and in my junior year, I began working with Professor Hima Lakkaraju at Harvard's AI4LIFE lab. After pursuing an independent study exploring the interactions between various explainability methods, as well as a study evaluating the fairness of race norming in lung capacity equations [8], **I, alongside my colleagues, built OpenXAI**. OpenXAI is an open-source framework published at NeurIPS 2022 [9] that evaluates and benchmarks the faithfulness, stability, and fairness of post-hoc explanation methods with an easy-to-use API. I built several aspects of the OpenXAI framework from the ground up, including the entire fairness pipeline and implementations of several new post-hoc gradient explanation methods. This experience harnessed my ability to both invent and test complex algorithms and metrics, and to build a large, constantly growing multi-contributor library.

During my second summer at IBM Research AI, discussions with NLP research teams and an exploration into language model auto-prompting sparked an interest in a new field: the potential of Large Language Models (LLMs), their generative capabilities, and their applications across social networks, healthcare, finance and beyond. Current LLMs have captured widespread attention largely due to their excellent sequence prediction capabilities. Language understanding, however, needs to involve *both* syntactic *and* semantic abilities. While LLMs are fantastic syntactic predictors (given just how broad their training corpus is), their statistical basis in the Transformer architecture [10] means they are unable to capture the abstract world-modeling and symbolic reasoning (esp. causality) that are crucial to real human language understanding. Until we address this shortcoming, LLMs will be unable to reach their goal of impacting practical societal applications at-scale. **In this vein, I am currently pursuing thesis research at the intersection of natural language processing, reasoning, and explainability** in Professor Lakkaraju's lab**.** My work includes improving the base question-answering ability of NLP models through an extension ofLM pre-training, where I augment training data with information from few-hop entity linking in large knowledge graphs like Wikidata. I accomplish this through a novel masked language prediction objective which encapsulates the efficacy of capturing causal and knowledge graph dependencies between entities for improved question answering abilities. Although there are many datasets for measuring the general logic abilities of language models through Q&A, there is a lack of data sources that explicitly focus on complex causal reasoning

and include high quality explanations of those answers as well. I have addressed this gap by building a **first-of-its-kind dataset called CReDET: Causal Reasoning Dataset and Explanation Testsuite** [11]. CReDET is carefully curated from questions in professional law, medicine, and management admission tests with the goal of measuring the accuracy of LMs for causal reasoning, along with their ability to explain their reasoning (high quality human annotated explanations provided for each causal reasoning Q&A pair). I am submitting this work to ICML 2023.

My experiences have solidified my passion for research into Generative AI, large language models, and reasoning and trustworthy AI. True natural language understanding abilities will have tremendous implications as we aim to build healthcare and clinical medicine applications, study and moderate the implications of social networks, and beyond. I envision a paradigm in which knowledge, reasoning, and interpretability all fuel each other in a seamless cycle enabling language models to hold advanced semantic ability along with reasoning, as well as the capability to explain their decisions. This is an inherently interdisciplinary goal, and I am excited to integrate lessons from Cognitive Science and Linguistics into my work.

In a PhD, I am excited to work towards this vision by exploring both NLP and Interpretable AI. I am always open to new, exciting research problems, but listed below are potential initial directions I envision exploring:

- **Improving Reasoning ability of LLMs** through knowledge graphs (KG) and their integration with LMs. There are many potential paths to explore this and I can envision mapping the KG relationships to propositional logic through their Conjunctive or Deterministic Decomposable Negation Normal Form, with embeddings. I could further explore learning vector based representations that can be naturally handled by neural networks - augmenting the training objective of LMs with an additional logic loss as a means of applying soft-constraints. This will enable the LM to incorporate knowledge between multi-hop KG paths with the objective of improving its reasoning ability.
- **Language Model - Explain your reasoning!** In my PhD I hope to explore the fertile intersection of language modeling, reasoning and explainability by addressing another key goal - developing methods to enable LMs to explain their reasoning answers. Infusing LMs with self-explanation ability will both allow for broader use and assist developers in improving the models themselves. I can envision automatically augmenting input prompts with relevant knowledge segments extracted from both causal graphs and general knowledge bases. I believe my significant research experience in explainable AI and inherently interpretable AI architectures [4] has solidly prepared me to further explore new architectures and attention mechanisms to solve the open research problem of allowing LMs to explain their reasoning.

My long term goal is to become a professor. A career in academia will allow me to both pursue cutting edge research *and* create strong teaching and mentorship systems to support young, historically underrepresented groups in the field. This is an effort that I am extremely passionate about and have worked towards as the Co-President of Harvard WiCS (Women in CS, Harvard's largest undergraduate student organization with 600+ members) and a lead Teaching Fellow for AM 120 - Applied Linear Algebra and Big Data (efforts for which I was awarded Harvard's Derek Bok Prize for Excellence in Teaching). As a Harvard Technology Innovation Fellow, I have learned just how crucial the intersection of technology and society is, and during my PhD, I hope to work towards the creation of robust and trustworthy AI innovations that benefit and interact positively with the world around us.

Achieving the end goal of human-level language understanding will require interdisciplinary insights from cognitive science, natural language processing and CS, linguistics, and psychology, and in this vein, **at MIT IDSS**, I am eager to work with Professors Robert Berwick, David Sontag, and Tommi Jaakkola. Having followed the publications of professors in IDSS, I am confident that there is a clear fit for my interests and skills, and I know that pursuing a PhD at MIT's IDSS will allow me to learn and thrive. I also deeply value MIT's commitment to DEI initiatives, and I look forward to getting involved with MIT's Graduate Women in Course 6 group (GW6) and the Thriving Stars program to continue my work with teaching, mentorship, and community-building.

References:

[1] **Isha Puri** and David Cox, "A System for Accurate Tracking and Video Recordings of Rodent Eye Movements using Convolutional Neural Networks for Biomedical Image Segmentation," 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2018, pp. 3590-3593.

[2] **Isha Puri**, "DYSCERN - A Scalable and Freely Accessible Machine Learning Based Web Application for the Early Detection of Dyslexia," NCWIT Collegiate National Award, 2021.

[3] K. Milton. Summation techniques, Padé approximants, and continued fractions. 2011.

[4] **Isha Puri**, Amit Dhurandhar, Tejaswini Pedapati, Kartikeyan Shanmugam, Dennis Wei, and Kush R. Varshney, 'CoFrNets: Interpretable Neural Architecture Inspired by Continued Fractions', in Advances in Neural Information Processing Systems, 2021, vol. 34, pp. 21668–21680.

[5] Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, Geoffrey Hinton. Neural Additive Models: Interpretable Machine Learning with Neural Nets. In *arXiv: 2004.13912,* 2020.

[6] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman and Hall/CRC, 1990.

[7] IBM AIX360, CoFrNets, https://github.com/Trusted-AI/AIX360.

[8] **Isha Puri**, Neil Sehgal, Usha Bhalla. 'Reconsidering the Algorithmic Fairness of Race Adjustment in Pulmonary Function Equations', AI for Social Good Workshop, AAAI Conference on Artificial Intelligence, 2023.

[9] Chirag Agarwal, Satyapriya Krishna, Eshika Saxena, Martin Pawelczyk, Nari Johnson, **Isha Puri**, Marinka Zitnik, and Himabindu Lakkaraju. 'OpenXAI: Towards a Transparent Evaluation of Model Explanations', in Advances in Neural Information Processing Systems, 2022.

[10] **Isha Puri**, and Himabindu Lakkaraju, "CReDET: Causal Reasoning Dataset and Explanation Testsuite", Submission to International Conference on Machine Learning, 2023.