**Thesis Project Portfolio**

# Recommendations for UVA CS New Curriculum Enhancement

(Technical Report)

# Synthetic Data, Generative Artificial Intelligence, and Mitigating Bias

(STS Research Paper)

An Undergraduate Thesis

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

**Samyak Thapa**

Spring, 2024

Department of Computer Science

**Table of Contents**

## Introduction

The technical report for my computer science capstone and my STS research paper are unrelated. My technical report focuses on my recommendations to improve the computer science department's new curriculum, from my perspective as part of the first cohort of students graduating with this curriculum change, and as a teaching assistant. My STS research paper focuses on recent developments regarding synthetic data, which is artificially generated data as a method for assisting the training of machine learning models. It encompasses this topic in the larger development of generative artificial intelligence, with perspectives from an ethical point of view. While my technical topic is certainly important to me as a graduate of the computer science program, I chose to pick an entirely new topic for my STS portion because the technology already has, and will continue to, profoundly impact the world.

## Technical Topic

The technical portion of my thesis produced a few recommendations for enhancements of UVA's computer science curriculum. Beginning in Fall 2021, UVA's computer science department rehauled its curriculum, to remove redundancies in course offerings and to align it with that of other schools. Since I will be graduating 2 semesters early, I am part of the first cohort of students with the new curriculum courses. As such, I wanted to provide my thoughts regarding the curriculum redesign. The main proposals are to reduce the focus on software engineering, adding topics from electrical and computer engineering, and once again increasing focus on algorithmic proof writing and correctness. This can be achieved by removing CS 3240 (Advanced Software Development Techniques / Software Engineering) as a requirement for the B.S. CS degree, adding more topics from digital logic design into the curriculum, notably those

missing from CS 2130 (Computer Systems and Organization 1), and removing the machine learning content from CS 3100 (Data Structures and Algorithms 2) for a focus on algorithm correctness instead. I believe these changes will increase student preparedness for lower-level CS roles, improve graduate school preparedness, and overall strengthen the competitiveness of UVA CS graduates, without sacrificing knowledge in software engineering.

## STS Topic

In my STS research, I explore the overall use of synthetic data as well as its promise as an aid for machine learning training processes. It is a relatively new topic of research and goes hand-in-hand with generative artificial intelligence as Generative AI is one way synthetic data is created. Synthetic data usage is becoming ubiquitous in the machine learning world, from being used to improve medical diagnosis models to account for edge cases and extremes, to being used by car companies to improve their autopilot systems. It can even be used to mitigate gender bias associated with text in large language models. However, synthetic data is also dangerous. It is central in the discussion about deepfakes, misinformation, and bias. Using synthetic data can perpetuate dangerous biases in models. With an infinite amount of any kind of data you want, the possibilities can be catastrophic. One possibility that comes to mind is any government creating facial recognition models tailored to a certain race, using synthetic images as its foundation. Another is the spread of misinformation online via AI generated articles and deepfake voices and videos, particularly in times where the truth matters most, like an election. This topic will become increasingly relevant in the coming years, and needs to be tackled from ethical, legal, and technological perspectives.

## Conclusion

Though my technical and STS topics differed significantly, researching my STS topic allowed me a better perspective on how to improve our CS curriculum. Considering how specialized research is in the AI and ML fields, it shows that an undergraduate computer science curriculum should be broad. This way, if students choose to pursue graduate studies, they have more flexibility in studying anything they wish.

Researching both curriculum design and more on generative AI has made me ponder what the future of education looks like with generative models like ChatGPT so readily available. An interesting perspective that would combine both topics is how computer science education needs to change in schools to account for the assistance AI can provide. Another could be studying how much access to generative AI would increase or decrease the gap between students who have constant access in developed countries versus students in others who do not. There are certainly interesting ethical questions to be considered from these topics.

## Acknowledgements

I'd like to acknowledge my STS professor Dr. Richard Jacques for his continuous support of my STS senior thesis in both STS 4500 and 4600. His feedback and assistance were invaluable.