**Thesis Project Portfolio**

**Solving Pay Stub Compliance Issues Through Asynchronous Generation and Persisting to S3**

(Technical Report)

**Investing the Effects of Algorithmic Bias in Lung Cancer Diagnosis and Treatment**

(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

**Preethi Chidambaram**

Spring, 2023

Department of Computer Science

## Table of Contents

Sociotechnical Synthesis

Solving Pay Stub Compliance Issues Through Asynchronous Generation and Persisting to S3

Investing the Effects of Algorithmic Bias in Lung Cancer Diagnosis and Treatment

Prospectus

## Sociotechnical Synthesis

Data mismanagement within technological products can have serious consequences because it can undermine the user's trust and potentially harm them depending on the type of data being managed. My technical and STS topics explore the nature of technical systems that have underlying flaws in their backend data models and how this can negatively impact user experience. My technical topic focuses on the refactoring of an automated payroll system's backend database while my STS topic examines the ethical implications of medical AI systems that are not trained on representative datasets. Both projects highlight that effective data management is a key step in the development and maintenance of technical systems.

Gusto is a company that provides payroll and employee management services to small businesses. The company's automated payroll platform does not keep track of employee pay stub version history because the pay stubs are generated on demand by aggregating employee data. This means that the pay stub file is not stored anywhere in the system, which can lead to compliance issues with industry standards if a user wants to access an overwritten file. To address this compliance issue, I helped restructure the backend database by creating a model to store pay stubs and persist them to S3, Amazon's cloud storage service. Adding this backend data model allowed me to improve the overall pay stub file generation process from being on demand to asynchronous. Thus, every time the system finished running payroll for an employer, it would automatically go ahead and generate all employee pay stubs for that company at one time and store them in Amazon S3. Storing the pay stub files in S3 proved to be a simple, cost-efficient solution to tracking version history because it didn't take up space on the payroll system itself, which maintained the performance of the system. Throughout the refactoring process, I

found that preserving aspects of the existing pay stub behavior was important for maintaining a satisfactory customer experience.

Data collection and training is the most important yet most tedious part of developing AI models. It is difficult to gather representative training data for medical AI systems because the quality and quantity of available data is generally limited, which can lead to inherent biases and lower accuracies within the resulting model. With regards to AI systems used to diagnose and treat lung cancer, there are a few well established public datasets, but a problem that comes with using these datasets is sampling bias and overfitting. Lung cancer diagnosis systems depend heavily on annotated images of a patient's lungs, but medical datasets are often less comprehensive than typical computer vision datasets and thus contributes to overall inaccuracies within the systems. One of the biggest existing disparities in the treatment of lung cancer is that patients from lower socioeconomic backgrounds lack access to advanced treatment facilities, which delays their diagnosis and limits the available treatment options. In analyzing this situation under Actor Network Theory, I found that health insurance companies and oncologists play a crucial role in determining coverage policies for medical AI systems, which can considerably affect the accessibility of these systems. While there is not much existing research on the relationship between insurance companies and lung cancer AI systems, my analysis demonstrates that oncologists are in the highest position of power within this network because they have to prove an AI system's clinical utility to insurance companies for them to determine the treatment's coverage options.

Working on both my technical and STS projects gave me a more holistic perspective on the significance of effective data management in technical systems and changed my mindset on the overall engineering design process. Both projects underscored the need for transparency and

accountability in the development and regulation of technical systems. My technical project demonstrated the effectiveness of industry standards because we were required to refactor the system due to compliance issues. On the other hand, my STS topic showed that the lack of standardization amongst medical AI devices can lead to problems with regards to system bias as well as accessibility. My technical project taught me the importance of balancing system performance with consumer needs, while my STS topic highlighted the ethical consequences of technical systems trained on biased or limited datasets. Combining these two perspectives has shown me that effective data collection and evaluation can lead to the development of more robust and responsible technical systems.