**Creating An Autonomous Laboratory Navigator**
(Technical Project)

**Decoding how Cognitive Bias Becomes Algorithmic Bias in Artificially Intelligent Decisioning Systems**
(STS Project)

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Cobrina Chiu

November 1, 2022

On my honor as a University student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments.

ADVISORS

Joshua Earle, Department of Engineering and Society

Tariq Iqbal, Department of Engineering Systems and Environment

## Introduction

Decision-making happens everywhere, from mundane chores to high-stakes court hearings. With the advent and rise of artificial intelligence (AI), it has become more common to see computers follow or even dictate the decisioning patterns of humans. In many ways, this algorithmic guidance has helped – think of a GPS calculating the most efficient path to a destination. Unfortunately, AI is not immune to the biases present in the society it is embedded in (Buolamwini, 2018), and in my STS research, it is my goal to understand what role human cognition plays in the production of algorithmic bias.

My technical project is one application of an AI decisioning problem. The objective is to create a functional robotic tour guide that can autonomously navigate a user through the University of Virginia's Link Lab. This technical project is a demonstration of the technology I will be exploring in my STS project: AI decisioning systems. Through this exploration, I will combine my backgrounds in computer and cognitive science to highlight the factors that cause human prejudice to manifest in AI-made decisions. I also will come to an understanding of the real harm algorithmic bias inflicts on marginalized communities and how engineers can create safeguards. The answers to these questions are important due to the ethical dilemmas presented by the rise of AI (Verbeek, 2019), especially regarding the preservation of human wellness. It is especially critical to debunk the idea that the decisions a computer makes are the optimal ones (O'Neil, 2016).

In the prospectus that follows, I will formally introduce both projects and outline the methods and foundational texts I have selected to guide my research.

## Technical Project

Anyone who has ever walked into the Link Lab at UVA understands that it is an intimidating space to navigate. The goal of this technical project will be to create a robotic assistant that guides a user from the entrance of the lab to the location within the lab they wish to go. In implementing this device, newcomers and visitors to the Link Lab will be able to traverse the space more efficiently.

This project will be implemented with a Ohmni robot, a telepresence robot often used in office spaces by employees that work remotely during the pandemic. It allows a person to connect to it over the Internet and remotely control the robot's movement while seeing and hearing what the robot can. Rather than using the video call function, the robot's display screen will run an interactive user interface prompting visitors to select a location to go to. Once a location is selected, the robot will move to that location while avoiding all obstacles. The robot then returns to its station at the front of the lab.

An add-on to this project will be to have the robot perceive that a person is standing in front of it and acknowledge the user verbally as a result. From then, the person would tell the robot where they want to go, and the robot would direct them on their path accordingly. The challenge of this task is robotic perception: how can we accurately get a robot to perceive when someone is requesting help? How will we overcome noise in the environment, both from passersby who do not wish to engage with the robot and from extraneous sounds?

With support from the Collaborative Robotics Lab, my goal is to complete this project in the fall semester and seek opportunities to conduct research with it in the spring.

**STS Project**

"Artificial intelligence" captures a wide variety of technologies (Wilson, 2010). To keep the scope of my research topic reasonable, this project will focus on those technologies involved in decision-making algorithms. These systems are often trained using machine learning (ML) to parse information from a given dataset and apply the knowledge it acquired to arrive at some conclusion about the task at hand (Soleimani, 2021). This process is made possible by a technique called deep learning or neural networks which takes inspiration from the neural connections of the human brain. Neural nets consist of many layers of connected nodes that process data and pass them forward between layers. To learn, the nets take input data and propagate it through the networks of layers and produce an output (Hardesty, 2017).

Because the technique behind AI is fundamentally math and computer driven, one might believe the decisions it makes must be objective. If this is the case, where does the bias come in? How is it possible that lines of code and numbers in a computer could exhibit the same logical flaws a human does? As AI becomes more ubiquitous in society, understanding the answers to these questions becomes more pressing for those involved in its work. Computers have embedded themselves in applications such as crime scene identification, medicinal dosage amounts, and job application tracking to name a few (O'Neil, 2016). Missteps in choices made in these situations can have dire consequences and multiplying this risk by the scope of AI's influence creates an ethical dilemma (Verbeek, 2019).

The impacts of AI on specific cultural groups have become more visible and researched in recent years. Black people, particularly Black women, are more likely to be inaccurately identified by facial recognition technology (Buolamwini, 2018). This has poor implications when combined with the rise of Big Data surveillance that sends police to disproportionally target

Black neighborhoods (O'Neil, 2016). In a lot of situations, any person with any distinguishing traits might be relevant to this project. When applying, demographic and other personal information one includes inadvertently casts them into buckets for the receiving end to filter through. As a result, people with disabilities, physical or mental health issues, women, and even anyone whose name is glaringly non-white are all disadvantaged by AI decision-makers (O'Neil, 2016). All these people are relevant, but it is too large of a population to consider when diving into this research topic, so I will narrow my social group to be non-white people.

The methods that will be employed in this STS project are finding, reading, and synthesizing previous literature and tracing networks of relationships. These methods are good for this research because of the volume of academic writings done on this topic, especially within the past three decades. This means more recent findings are widely accessible and can be directly applied to AI as it stands today. Tracing the networks of relationships is also intuitive for this research because the goal is to understand how human society factors into algorithmic bias, and part of that understanding requires analyzing the networks around AI decisioning.

Additionally, the STS frameworks used in this project will be case study and Black/race critique. The former will be the primary framework used in guiding this research; AI decisioning algorithms are common, and academia is teeming with literature detailing their consequences. The latter framework is an important lens to consider because while one social group might be most affected by problematic AI biases, folks from that group might not be the ones conducting this research and writing the articles I consider. Framing research in this topic around the criticisms of minority voices will better my understanding of how this technology interacts with race.

## Key Texts

The first primary source used for this STS topic will be Affect and Artificial Intelligence by Elizabeth Wilson. This book investigates the early days of AI and argues that newer developments in computation, namely the entry of cognition and minds, can be traced back to the origins of the science. Because my research topic is at the intersection of cognitive and computing science, I will be using this book as a basis for my understanding.

Joy Buolamwini's work on the disparities seen in commercial gender classification systems highlights a key example of widely used AI that fails to be unbiased. Her research article "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification" argues that several commercially used classification algorithms are more error-prone when classifying darker-skinned individuals, especially women. This is an important argument to consider when thinking about the implications of failures of AI, and I hope to use more of her writings about this topic as well to understand it through the lens of a Black woman.

"A Neural Network Framework for Cognitive Bias" by Johan Koteling, Alexander Toet, and Anne-Marie Brouwer argues that "many cognitive biases arise from intrinsic brain mechanisms that are fundamental for the working of biological neural networks" (Korteling, 2018, p. 1). This argument is relevant to my research topic as it establishes a link between the psychological processes in the human brain and the computational neural networks used in AI.

Finally, *Weapons of Math Destruction* by Cathy O'Neil details several instances of how reliance on numbers and data creates the tendency to enter feedback loops that inflict harm. O'Neil's book discusses the influence of Big Data on the criminal justice system, a topic I will likely address in my research and therefore find relevant to my topic.

## Works Cited


Buolamwini, J. & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. Retrieved October 12, 2022, from https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Hardesty, L. (2017, April 14). *Explained: Neural networks*. MIT News. Retrieved October 12, 2022, from https://news.mit.edu/2017/explained-neural-networks-deep-learning-0414

Korteling, J. E., Brouwer, A.-M., & Toet, A. (2018). A neural network framework for cognitive bias. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.01561

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Soleimani, Melika & Intezari, Ali & Taskin, Nazim & Pauleen, David. (2021). Cognitive biases in developing biased Artificial Intelligence recruitment system.

Verbeek, P. & Parizeau, M. (2019). Preliminary Study on the Ethics of Artificial Intelligence. Retrieved October 12, 2022, from https://unesdoc.unesco.org/ark:/48223/pf0000367823

Wilson, E. (2010). Affect and Artificial Intelligence. University of Washington Press Seattle, WA