The Cupola
Scholarship at Gettysburg College

CAFE Symposium 2024 Celebration of Academic First-Year Engagement

2-7-2024

Emergent AI

Jillian A. Bick
Gettysburg College, bickji01@gettysburg.edu

Emergent AI

Abstract

For many years, artificial intelligence (AI) was considered to be limited in its abilities due to being confined to a pre-defined set of data. Currently, however, AI models have grown in complexity and size, leading to some previously impossible behaviors. These behaviors, known as "emergent AI behaviors," are unpredictable and not pre-programmed. Their existence suggests that AI is expanding in adaptability and may one day rival human intelligence. Media often portrays AI as having emotions and having the ability to operate autonomously, but what behaviors are AI really capable of?

Keywords

Emergent AI, Artificial Intelligence (AI), Generative Agents

Disciplines

Artificial Intelligence and Robotics | Technology and Innovation

Comments

This poster was created based on work completed for FYS-179-2: How Much of Science Fiction is Science Fact, and presented as a part of the ninth annual CAFE Symposium on February 7, 2024.

Creative Commons License

Emergent AI

Jillian Bick

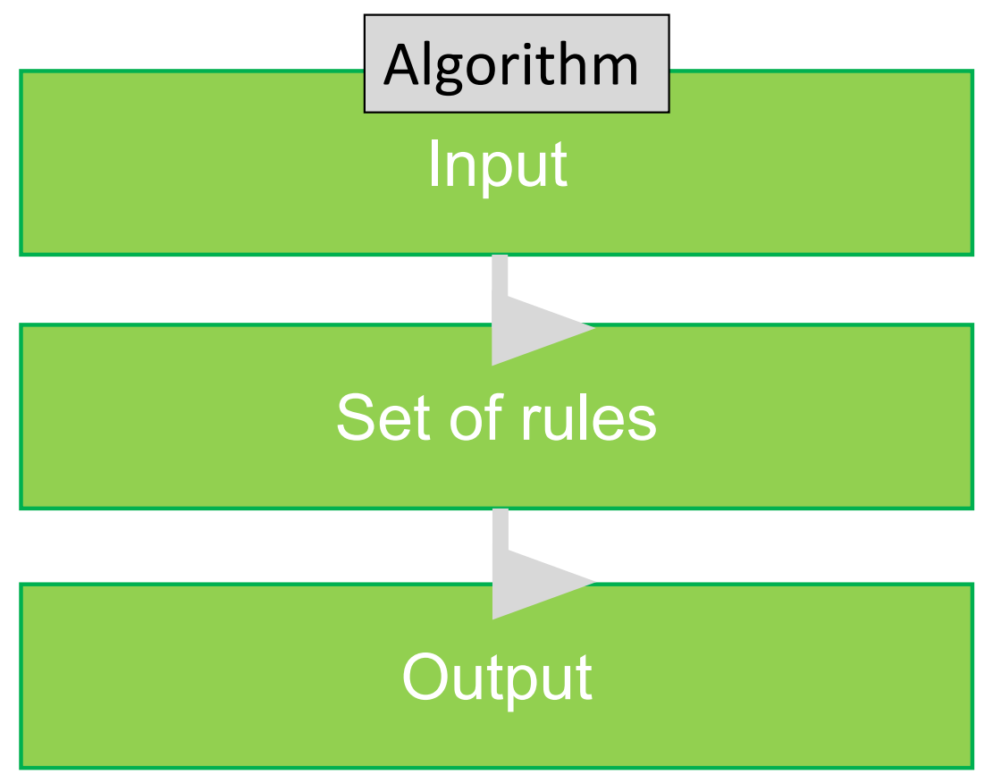
FYS-179-2 How Much of Science Fiction is Science Fact

Introduction

Artificial intelligence (AI) refers to a “human-made tool that emulates the “cognitive” abilities of the natural intelligence of human minds” to perform certain tasks (Tai, 2020). These tasks include problem-solving, pattern recognition, and the ability to make complex decisions. Current AI that can perform these tasks are classified as “narrow AI,” or AI that can only perform one task from a pre-defined set of data (Ornes 2023). In recent years, researchers have attempted to improve AI adaptability to new experiences and effectively mimic human intelligence despite the limitations of narrow AI. These developments have prompted discussions wondering how narrow AI truly is and the discovery of some unexpected behaviors. Media often portrays AI as having emotions and having the ability to operate autonomously, but what behaviors are AI really capable of?

How does AI Work?

- In the field of AI, there are currently two primary approaches for developing AI models:
- The first approach involves the use of an **algorithm**, or a set of rules for solving a particular problem. These algorithms follow a precise, step-by-step procedure, consistently producing the same responses to a given situation.
- The second approach is driven by data input into **artificial neural networks** inspired by the human brain. These neural networks receive programmed inputs of data that are subsequently translated into an intelligent output response.
- Both approaches are limited to a pre-defined set of data, meaning AI lacks creativity and common sense.

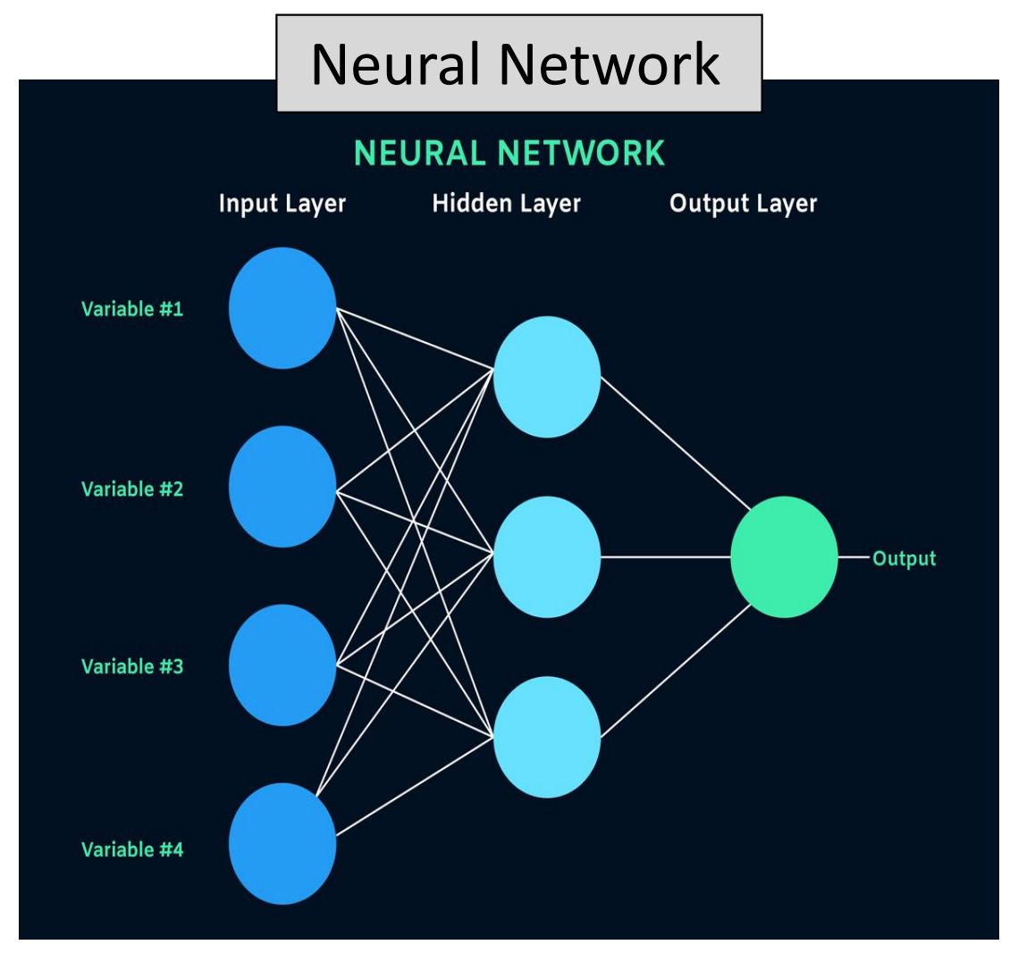


Algorithm

Input

Set of rules

Output



Neural Network

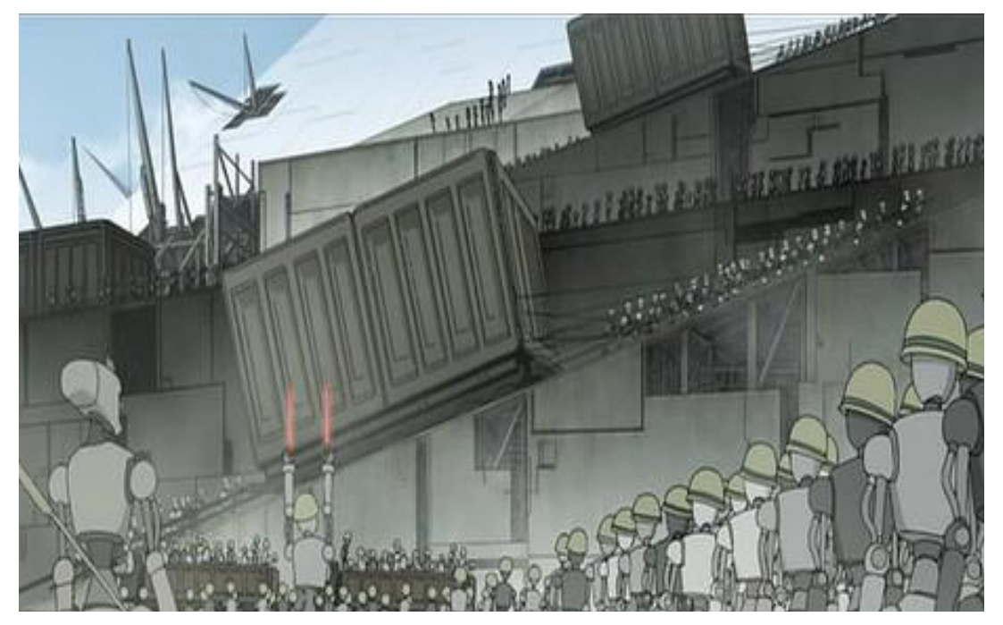
What is Emergent AI?

- **Emergent AI behaviors** are defined as behaviors that are not “not present in smaller [AI] models but [are] present in larger models” (Wei et al., 2022).
- Emergent AI behaviors have several defining characteristics:
  - Behaviors increase as AI complexity increases
  - Behaviors are not pre-programmed
  - Behaviors are unpredictable
- Research suggests emergent AI behavior is a direct result of AI interaction with other AI and the environment.
  - This is called the **Complex Systems Theory,** which characterizes AI models as “dynamic systems able to adapt in and evolve with a changing environment” (Chan, 2001).
    - In this way, AI are like ecosystems with multiple interconnected parts which interact and cause the entire system to undergo adaptations.
- Examples of emergent AI behavior include the ability to perform step-by-step reasoning and respond to the external environment with a logical plan or action.

Emergent AI in Media

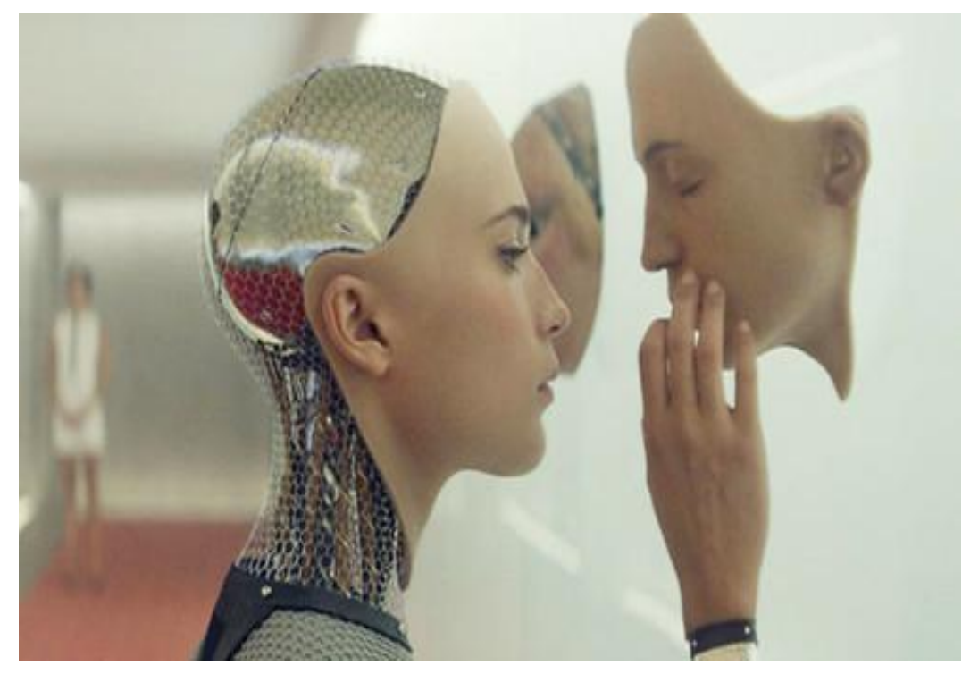
- In the video game *Detroit: Become Human*, androids are used as caretakers and workers for humanity.
- The story focuses on three protagonists, Markus, Kara, and Connor, who all exhibit emergent AI behavior by showing emotions and going against their pre-defined directives.
  - For example, Connor is a criminal investigator who is designed to catch “deviant” robots. Throughout the game, Connor can choose whether to follow this directive or not.

“Detroit: Become Human” by Quantic Dream. Public Domain.



"Animatrix" by Ed Pichler is licensed under CC BY-SA 2.0.

- In the animated series *Animatrix*, robots work perilous and service-based jobs for humanity.
- These robots exhibit emergent AI behavior after the AI B166ER kills his abusive human master and is executed himself. This sparks the robot rebellion which would eventually lead to the events of *The Matrix*.

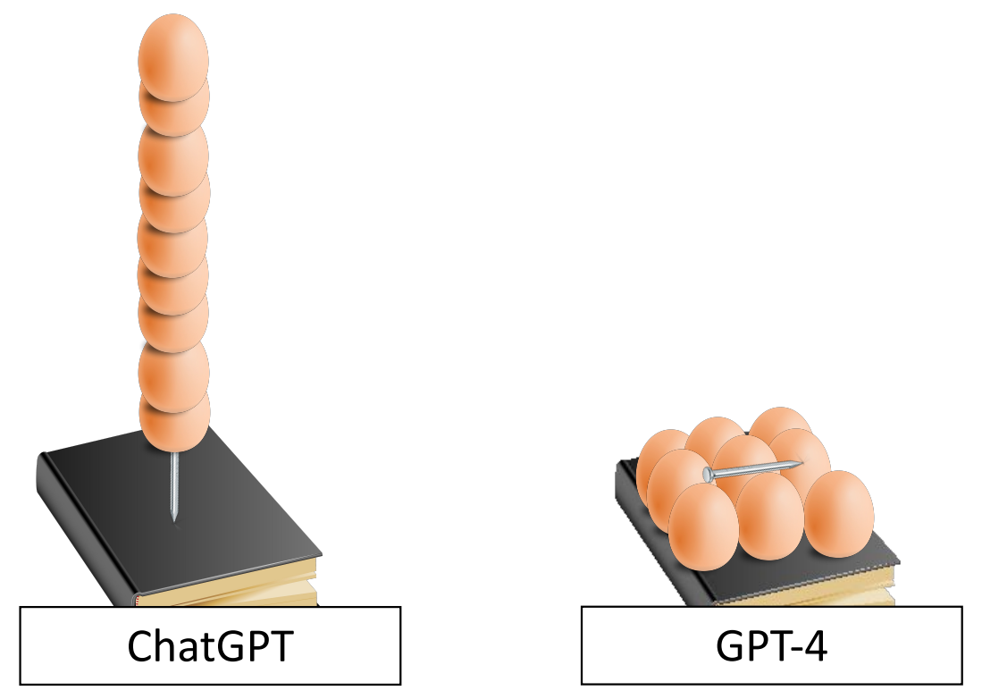


"ex-machina-movie" by Kanijoman is licensed under CC BY 2.0.

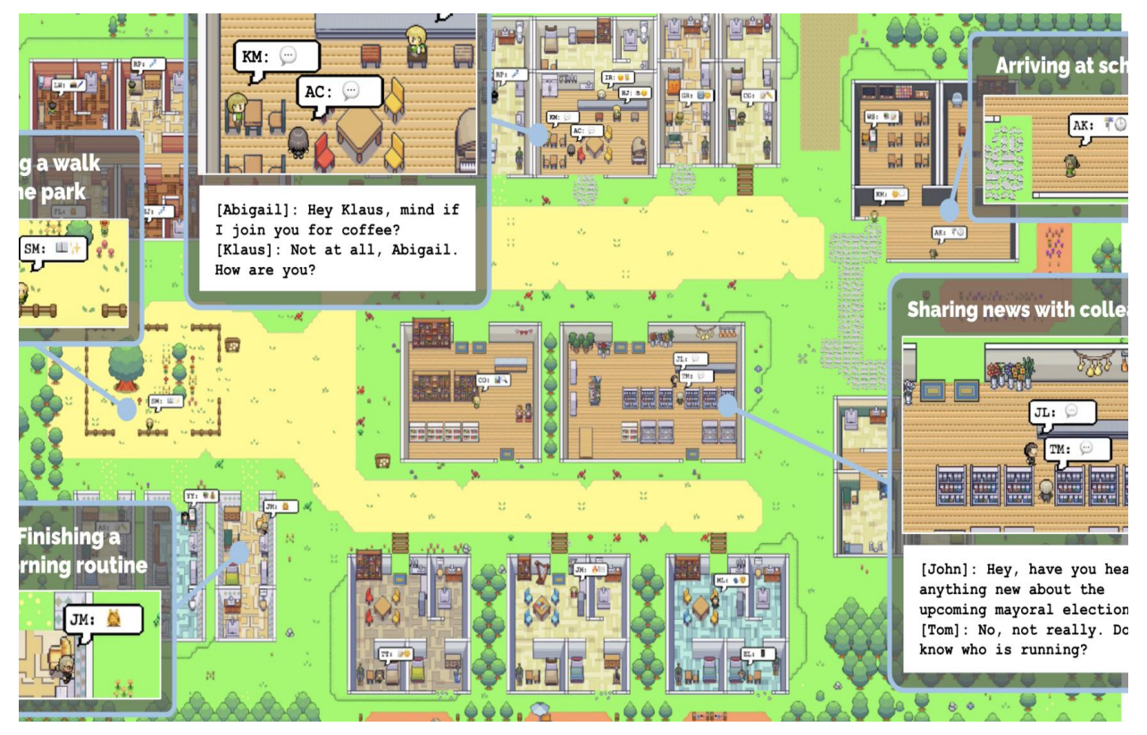
- In the movie *Ex Machina,* the protagonist, Nathan, interacts with the human-like AI Ava.
- At first, Ava seems to exhibit emergent AI behavior as she falls in love with Nathan; however, it is revealed that she had the primary directive of escaping the facility.
- She maintains a cold and detached demeanor despite interacting with Nathan, manipulating him to achieve her goal.

Reality of Emergent AI

- Sébastian Bubeck’s study “Sparks of Artificial General Intelligence: Early experiments with GPT-4,” presented two AI with the prompt of stacking a book, 9 eggs, a laptop, a bottle and a nail.
- A lack of common sense was considered one of the most fundamental flaws of AI; however, this study showcases that AI can develop it.

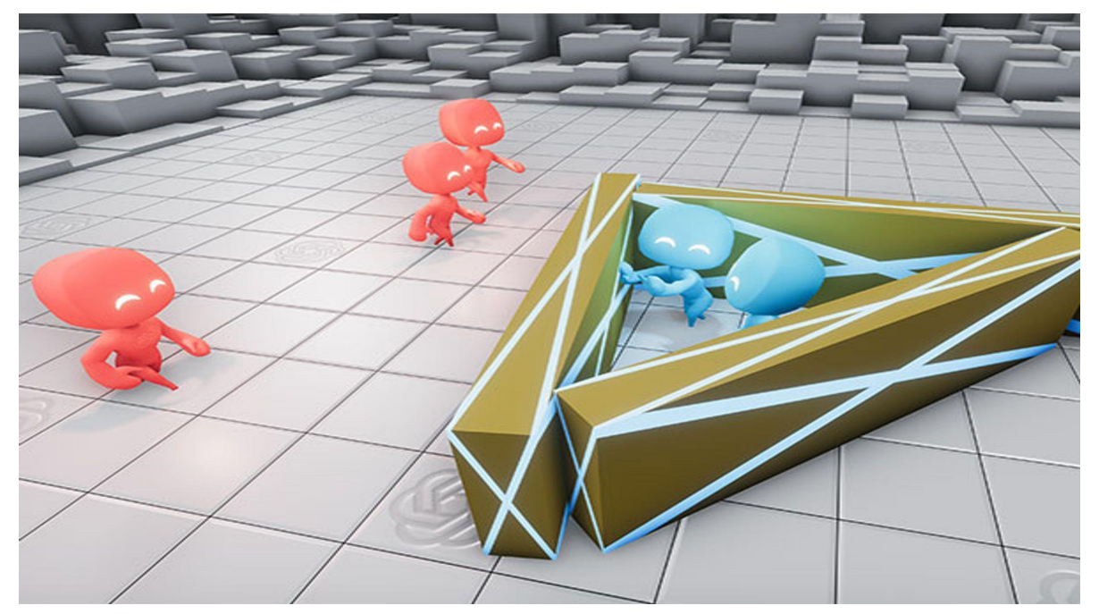


- Joon Sung Park’s study “Generative Agents: Interactive Simulacra of Human Behavior,” focuses on twenty-five interacting AI agents in a simulated town.
- All agents were pre-programmed with a base architecture simulating the processes of human thinking, while only one agent was assigned the primary directive of planning a Valentine’s Day party. They exhibited behaviors such as:
  - Decorating and inviting other agents to a Valentine’s Day party
  - Responding to situations such as a stove fire
  - Planning a schedule based on their hobbies or career



This image is from Joon Sung Park’s study “Generative Agents: Interactive Simulacra of Human Behavior.”

- OpenAI’s study “Multi-Agent Hide-and-Seek” had two AI teams compete for an award by playing hide-and-seek in a simple 3D environment.
- Their only incentive was to win hide-and seek.
- AI teams gradually developed strategies to achieve their goal of winning, using ramps and building fortresses.
- AI seekers even exploited a flaw in the environment.



This image is from Open AI’s study “Multi-Agent Hide-and-Seek.

Conclusion

While science fiction media often portrays AI as having complex emotions and autonomy, current studies on emergent AI behavior reveal that AI is not capable of completely replicating human behavior and intelligence. Nonetheless, these studies highlight the fact that the emergence of unpredictable and adaptive AI behavior shares similarities with the evolutionary adaptation observed in organisms as they grow and develop over time. This implies that continued exposure to simulations of real-life conditions as well as the creation of more complex AI models can foster the enhancement of emergent AI behaviors. Therefore, speculative portrayals in science fiction and real-world experiments together can provide a glimpse into the future of AI where models may one surpass initial limitations and expectations.

Works Cited

Bubeck, S., et al. (2023). *Sparks of Artificial General Intelligence: Early experiments with GPT-4*. https://arxiv.org/pdf/2303.12712.pdf

Chan, S. (2001). *Complex Adaptive Systems*. [PDF]. http://web.mit.edu/esd.83/www/notebook/Complex%20Adaptive%20Systems.pdf

Ornes, S. (2023, March 16). *The Unpredictable Abilities Emerging From Large AI Models*. Quanta Magazine. https://www.quantamagazine.org/the-unpredictable-abilities-emerging-from-large-ai-models-20230316/

Park, J. S. et al. (2023). Generative Agents: Interactive Simulacra of Human Behavior. *ArXiv:2304.03442 [Cs]*. https://arxiv.org/abs/2304.03442

Strickland, E. (2019, September 17). *AI Agents Startle Researchers With Unexpected Hide-and-Seek Strategies - IEEE Spectrum*. Spectrum.ieee.org. https://spectrum.ieee.org/ai-agents-startle-researchers-with-unexpected-strategies-in-hideandseek

Tai, M. C. (2020). The impact of artificial intelligence on human society and bioethics. *Tzu Chi Medical Journal*, *32*(4), 339–343. https://doi.org/10.4103/tcmj.tcmj_71_20

Tech Team. (n.d.). [Chart, How do Neural Networks Work?] Retrieved December 5, 2023, from https://blog.trustedtechteam.com/2018-05-16-artificial-intelligence-and-machine-learning-terms-you-need-to-know/

Wei, J. et al. (2022). Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*. https://openreview.net/forum?id=yzkSU5zdwD

Acknowledgements

I am grateful to Professor Ryan Johnson for making this project possible by nominating me through his seminar, FYS-179-2: How Much of Science Fiction is Science Fact.

I also want to thank my family for their unwavering support and encouragement throughout this journey. Special appreciation goes to my sister, Madison, for providing valuable feedback on this poster.