Developing an Interpretable Gray Box AI From a Game Theory Perspective

Artificial Intelligence (AI) is widespread in its uses, but there is an important flaw that lies unaddressed in the field. AI does not provide any reasons or give humans any insight in to its decision making process. For example, there have been many instances in high stakes situations where biases in the AI process have negatively influenced court trials and other high-stakes scenarios.

White Box AI vs Black Box AI

Artificial intelligence (AI) has proved its usefulness in countless fields such as medicine, criminal justice, finance. While AI is very effective, the most accurate AI models are also the most difficult to trust (Adadi & Berrada, 2018). If any human were to look inside an AI's decision process, they would only see countless seemingly random mathematical operations. However, the mysterious process allows AI to accomplish an enormous range of tasks, from translating paragraphs to identifying illnesses. The process AI takes to arrive at its conclusions is obscured beneath layers of seemingly meaningless calculations. AI produces exceptionally accurate results, but we do not know how the calculations relate to the inputs. These types of AI are called **black box AI** because their decision process is incomprehensible to humans.

However, AI was not always black box. The simplest and earliest forms of AI had parameters that related to the inputs and thus were very interpretable. The decision-making process of a white box AI gives insight into how different factors are weighted and why the AI decided upon a certain outcome. There are

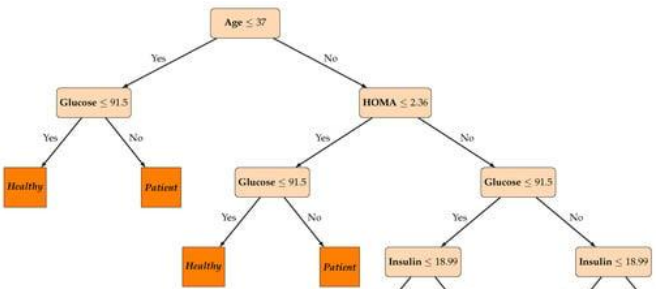

Figure 1: A decision tree, which is the simplest form of white box AI (Pintelas et al., 2020).

some types of AI (termed **white box AI**) that give us clear insights into their decision process. The simplest example of a white box AI is a decision tree (Figure 1).

The main drawback of white box AI is that it is not as accurate as black box AI. One study showed an approximately 10% difference between the accuracy of a black box AI and a white box AI across multiple datasets (Pintelas et al., 2020). White box AI is explainable; black box AI is accurate. However, both of these qualities are important to have in many cases.

Actor Critic Model

The actor critic model of AI involves divides the work of an AI agent into two parts. One part is called the actor, which takes in inputs from the environment and outputs

Current Areas of Research

The application of AI to scenarios where lives are at play has ethical concerns, especially if the AI is not interpretable (von Eschenbach, 2021). For example, AI has been used to predict the likelihood of a repeat offense for previously convicted criminals in trials. While it is very accurate at predicting a judge's decision, we do not understand the reasoning it takes. For example, it may be predicting a higher likelihood of a repeat offense based on factors such as race and income, copying bias in the training data. Any AI model can only be as good as its training data (Figure 2). Furthermore, it is very difficult to find the source of a bias when the AI system itself is incomprehensible (von Eschenbach, 2021). While high interpretability is not always called for, there is a need for both interpretable and accurate AI in high-stakes scenarios.

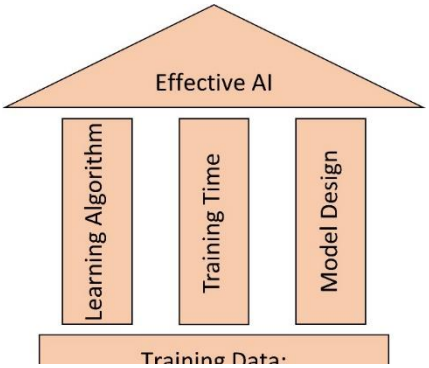

Figure 2 An inforgraphic about what makes an effective AI

Previous attempts have mainly focused on making black box AI more interpretable. This approach often involves using another AI to explain to humans what inputs are related to which outputs. Algorithms with the goal of explaining black box AI are called explainable artificial intelligence (XAI). Many XAI implementations that are currently used in industry are post hoc, meaning that fabricate an explanation based on the inputs and outputs. While somewhat effective, this method usually ignores the framework of the AI. Usually, this approach (termed model agnostic) does not make an attempt to understand what the AI is actually doing. It is equivalent to asking a doctor why a different doctor came to a certain diagnosis.

Another promising area of research is creating a gray box AI – combining both a white box and black box AI. The hope is to achieve a model with complete interpretability of a white box AI while having a black box AI to boost the accuracy. The black box AI generates fake training data based on its traning which is then used to further train the white box AI. However, there always seems to be a tradeoff between interpretability and accuracy (Pintelas et al., 2020). Data collected shows that gray box AI sometimes performs worse than black box AI (Pintelas et al., 2020). *Insert flowchart of how gray box AI uses black and white box AI*

It is clear that there is some factor that does not allow white box AI to be as accurate as black box AI. This factor is not accounted for in our current understanding of AI models. Even when the black box AI creates data for the white box AI to learn from, it is has difficulty reproducing the accuracy of a black box AI. There is a big difference in accuracy between black box and ensemble gray box models, but there is very little research on why it is not able to perform as well. The gap between black box and gray box AI in accuracy has simply been accepted as a truth by many research papers without understanding the underlying mechanism

driving the difference in accuracy.

With this research, I hope to start answering the question of where this difference originates from. I will be using a computer simulation to test how well various types of AI can communicate with each other in the context of a multiplayer team game.