Abstract

AI has become a ubiquitous approach to solving complex problems, but AI offers no explanation as to why it makes the decisions that it does. It was reported that less than 50% of individuals trust AI, and only 44% globally believe that the use of AI will have a positive impact. Previous research has observed a tradeoff between interpretability and accuracy when creating a model. This research explores the reasons why gray box AI may not be able to provide the same levels of accuracy as black box AI. I hypothesize that a prominent factor decreasing a gray box AI's accuracy is ineffective communication. I tested black box and white box agents in the simple crypto game from the Python library PettingZoo, in which agents must communicate in order to succeed. While the accuracy of the models was similar, the learning rates differed significantly. The black box - black box and the white box - white box team both converged to ideal play two times faster than the white box - black box team. These results demonstrate that there is a significant communication barrier between white box and black box AI, and this slows the learning speed of ensemble gray box models. This study provides a place to start from in investigating the relationship between interpretability and accuracy in AI models. Further research will involve exploring the causes of the difference between white box and black box AI or constructing AI models based on the information gathered in this study.

Keywords: artificial intelligence, game theory, interpretability, transparency, black box

The abstract would contain an overall summary of what you (as the author) would like to convey in the order of a general scientific article (Hook/Need, Background/Knowledge Gap, Methods, Results, Discussion/Conclusion, Applications). It would include some of the knowledge gaps that would eventually lead to researchable questions you have identified in the

field. You can build on the Abstract you have written for the Literature Review but reflect your experimental, procedure/prototyping, etc. Please follow the overall outline reviewed in class

Keywords: emotion understanding, interest, social development, prosocial behavior, infants

Acknowledgements

Thank you to my parents for supporting me and to Dr. Crowthers and Mr. Medeiros for providing guidance to my project.