Thesis Portfolio

Towards Transparent Robotic Planning via Contrastive Explanations
(Technical Report)

Safety Issues of Black Box Models in Autonomous Systems
(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

Shenghui Chen
Fall, 2020

Department of Computer Science

Table of Contents

Sociotechnical Synthesis

In recent years, dramatic success in machine learning has led to a torrent of Artificial Intelligence (AI) applications. Continued advances promise to produce autonomous systems that will perceive, learn, decide, and act on their own. Notable examples include autonomous driving, drones, medical assistive technologies. However, as these applications show great potential for improving the quality of life and more convenience for mankind, one also has to recognize the risks it brings. In particular, the commonly used deep learning modules in such autonomous systems are widely believed to be powerful despite inherently uninterpretable and complicated, making them black box models. In this thesis portfolio, I have conducted research on the technical front on improving the transparency of robotic planning systems through the generation of contrastive explanations, which was published in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020*. From a Science, Technology and Society (STS) perspective, I analyze the status quo, possible causes, and impact on different stakeholders of black box models in safety-critical autonomous systems, and demonstrated my argument through two case studies.

Both technical and STS approaches are employed in this thesis as I believe understanding the societal factors in any technical topic is as important as finding a technical fix because it reveals the full picture and potentially clarify the root of the problem, which in many cases needs collaboration from different sectors of the society to solve.

In this work, the technical paper delves deeper into the domain of Explainable AI (XAI), aiming at improving user understanding and trust of autonomous robotic planning systems by introducing new concepts and algorithms to generate contrastive explanations. The STS research

provides the broader context of the technical work, outlining the problems and potential solutions of safety issues of black box models in autonomous systems.

The technical portion of my thesis first recognizes that providing explanations of chosen robotic actions can help to increase the transparency of robotic planning and improve users' trust. Also, we learn from literature survey of social science works that the best explanations are contrastive, explaining not just why one action is taken, but why one action is taken instead of another. We formalize the notion of contrastive explanations for robotic planning policies based on Markov decision processes, drawing on insights from the social sciences. We present methods for the automated generation of contrastive explanations with three key factors: selectiveness, constrictiveness, and responsibility. The results of a user study with 100 participants on the Amazon Mechanical Turk platform show that our generated contrastive explanations can help to increase users' understanding and trust of robotic planning policies while reducing users' cognitive burden.

In my STS research, I first introduce the status quo of the safety issues underlying in black box models of current autonomous systems. I apply the STS framework of *Technological Momentum* proposed by Thomas P. Hughes, arguing we are still in the first phase of deploying autonomous systems and our society still has the ability to steer the development towards a more safety-focused, privacy-oriented direction. Then I exemplify the problems of *technically black box* and *socially black box* models through case studies of autonomous driving bugs and Boeing 747 crashes. Finally, I analyze the stakeholders involved and their corresponding responsibilities on this issue.

I believe the technical and STS research in this thesis complement each other, together contributing to the theoretical understanding and practical progress of more transparent and

trustworthy autonomous systems. Many problems discussed in the thesis have deep ethical significance for everyone involved in technology to consider, for example how should corporations balance the commercial benefits and potential risks of employing black box models in autonomous systems.

I started thinking and investigating in this area since my third-year, and the more I learn about the problems remain, the more I am excited about it. I am attracted not only by the intellectual challenges posed by the issues, but more importantly because I realize the significant impact autonomous systems will have on everyone's daily lives. From the experience of researching and writing this thesis, if I learned anything more critical than the project itself, it is power of passion in a subject. It is common to encounter obstacles of all kinds in research, but genuine interests in something can bring you out of failures and carry you on. This is the only advice I can give to later SEAS students: find a domain you truly love and do your research with hard work.

I would like to express my appreciation to my STS advisors Professor Toluwalogo Odumosu, Professor Richard Jacques, and my technical advisor Professor Lu Feng. In particular, I want to thank Professor Lu Feng for her patient advice and guidance since the beginning of my research journey. I also want to thank my friend and research partner Kayla Boggess for the enjoyable collaboration we had in my technical project. Last but not least, I would like to thank my family and friends for their support, encouragement, and constant source of inspiration.