Towards Transparent Robotic Planning via Contrastive Explanations

(Technical Paper)

"Efficiency Versus Transparency" of Black-box Models

(STS Paper)

A **Thesis Prospectus Submitted to the**

Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements of the Degree
Bachelor of Science, School of Engineering

Shenghui Chen
Fall, 2020

Technical Project Team Members
Shenghui Chen
Kayla Boggess
Lu Feng

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

Signature Shenghui Chen Date 2020/10/19
Shenghui Chen

Approved Lu Feng Date Oct 19, 2020
Lu Feng, Department of Computer Science

Approved Date 10/20/2020
Toluwalogo B. Odumosu, Department of Engineering and Society

A. Introduction

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 potentials for improving the quality of life and more convenience for mankind, one also has to recognize the risks it brings. An important factor that limits the effectiveness of these systems is the machine's current inability to explain their decisions and actions to human users, which can lead to insufficient understanding and inappropriate trust from the users. My technical project looks into one potential direction to this problem: Explainable AI (XAI), where aims at creating more explainable AI models and enable human users to understand, appropriately trust, and effectively manage the intelligent agents (Turek, 2016). In particular, my research group formalizes contrastive explanations for MDPs based on three key factors, implements a prototype to automatically generate contrastive explanations, and conducted a user study with 100 participants to validate the effectiveness of the designed system.

For the STS project, I will examine the dilemma of efficiency versus transparency from a historical perspective, tracing through the origin of thoughts of AI community and relate it with the period of empiricism versus rationalism. Using the technology momentum as the theoretical framework, I will attempt to build an argument for the explanation of why, where and how the conflicts between two AI directions, efficiency and transparency, emerge.

B. Technical Project

With the rise of Artificial Intelligence and Machine Learning techniques, people are increasingly aware of the need for enhancing the transparency of AI decision-making systems in order to improve users' trust. Because of traditional "black-box" approaches, lay-users have little understanding of how a decision is made or why an action occurs, often leading to misunderstanding and mistrust of the system, which can further lead to problems caused by system misuse (Goebel et al., 2018). The vast majority of work in this direction has been focused on the building of simplified interpretable models as approximations of complex decision-making functions (Mittelstadt et al., 2019). However, few works consider social science theories or explanation. For example, Miller suggests humans prefer contrastive explanations, or explanations that revolve around counterfactuals (Miller, 2017). Specifically, humans tend to ask not why an event P happens, but why an event P happens instead of some event Q. Understanding this contrast of events is more important to the human user than statements of probabilities or lists of total causes. In this project, in order to concretize the argument, we consider a motivating example which is a route planning task for a robot navigating in a grid map shown in Figure 1.

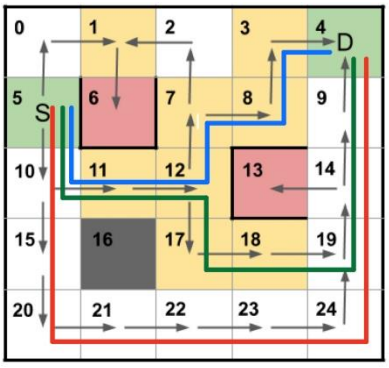

Figure 1 An example grid map for robotic planning: Green grids: start (S) and destination (D). Black grids: buildings. Red grids: dead-ends. Yellow grids: urban roads. White grids: highways.

Inspired by the insights from the social science theories, my technical project formalizes the notion of "contrastive explanations" in the context of robotic planning based on Markov

decision processes (MDPs), which is a popular modeling formalism for representing abstract robotic mission plans (Thrun, 2000). Our goal is to explain action choices in a planned robotic route, which can be computed as the optimal MDP policies using reinforcement learning (Sutton and Barto, 2018) or formal methods (Lacerda et al., 2019). More specifically, we focus on three key factors of contrastive explanations: selectiveness (e.g., choosing the most relevant events), constrictiveness (e.g., numbering how many future possible actions than an action causes) and responsibility (e.g., rating how important an action is in causing an event). We first formalize three factors mathematically in the context of MDPs. Then, we propose separate algorithms to compute the three factors in order to generate the contrastive explanations. Finally, combining a language template and the results of the three factors, we present a python program to automatically encode the mathematical results into a human-understandable contrastive explanation.

Explanation Type	**Explanation Example Text**
No Explanation	N/A
Naive Explanation (One State)	We move east at grid 10.
Responsibility Explanation	We move east at grid 10 because it leads to the shortest route.
Constrictive Explanation	We move east at grid 10 because it leads to the most flexible future route.
Naive Explanation (Entire Path)	First, we move south at grid 5. Next, we move east at grid 10. Then, we move east at grid 11. Next, we move north at grid 12. Then, we move east at grid 7. Next, we move north at grid 8. Finally, we move east at grid 3.
Selective Explanation	First, we move south at critical grid 5. Then, we move east at critical grid 10. Next, we move north at critical grid 12. Finally, we move east at critical grid 7. All other decisions result in equivalent routes.
Contrastive Explanation (All Factors)	First, we move south at critical grid 5 because it leads to the shortest and most flexible future route. Then, we move east at critical grid 10 because it leads to the shortest and most flexible future route. Next, we move north at critical grid 12 because it leads to the shortest route. Finally, we move east at critical grid 7 because it leads to the shortest route. All other decisions result in equivalent routes.

Table 1: Explanations generated by the algorithm for the running example in Figure 1

To assess the effectiveness of the approach proposed above, we have conducted a user study. For this study, we recruited 100 individuals using Amazon Mechanical Turk. We asked them to evaluate different types of explanations in Table 1. Users were presented with 3 different 10-by-10 grid maps, each containing an optimal route from a start state to a finish state. Each route was presented with explanations about the robotic actions taken within it of 7 different types: no

explanation, naïve explanation on one state, responsibility explanation, constrictiveness explanation, naïve explanation on the whole path, selective explanation, and contrastive explanation. Then, we measure the user understanding and user trust using a 5-point Likert scale, and users were also asked to choose the preferred explanation out of different groupings of explanations. We also measured time spent accessing the explanation as an objective dependent variable.

We hypothesize that contrastive explanations will 1) increase user understanding of the information given, 2) enhance user trust in the correctness of the autonomous system, and 3) users will prefer contrastive explanation over other types of naïve explanation.

The results show that, 1) the use of responsibility and constrictiveness increase user understanding, while selectiveness slightly decrease user understanding. This may be justified by the user's perception that more information is better, though this may not be the case factually. Overall, contrastive explanation increase understanding and decrease cognitive burden. 2), the use of responsibility increases user trust, while the use of selectiveness decreases this factor. Constrictiveness has no effect. Overall, contrastive explanations increase user trust using responsibility justification. 3), users prefer responsibility and constrictiveness explanations over naive explanations, but do not prefer selective explanations. Users prefer contrastive explanations at the same rate as single factor explanations.

From the user study, we understand that our proposed method and generated explanations are mostly effective at increasing user understanding and enhancing user trust, but a more thorough study could further reveal the reasoning behind some unexpected user behaviors. Furthermore, another important direction ahead is to assess the computational cost of generating such

explanations and whether managers of these autonomous systems are willing to sacrifice some performance of the system (e.g. speed, capacity, etc.) in order to achieve a higher level of transparency. This is a great unknown in the future, and directly related to the STS research proposed below.

C. STS Project

My STS research project is a direct extension of my technical project, in a sense, I am taking this opportunity to investigate the same problem from a different perspective at a broader scale. Specifically, I am still interested in the tradeoff between efficiency and transparency of a black-box system. Rather than delving into one of the possible solutions like I did in the technical project, I want to take a step back to fully understand the origin of this problem from a historical perspective using the STS theoretical frameworks we learned in class.

Since the success of ImageNet and AlphaGo, AI and machine learning techniques like the various kinds of neural networks and reinforcement learning has been a hotspot on the news and in social discussions. People now often have the impression that these systems are black-box models and are too complicated for regular users to fully grasp with. It is only natural that some people are distrustful of such unknown and go down the sci-fi path of fearing for singularities. However, if we put things in a broader temporal scale, one will see how the technology evolved to this day and have a deeper understanding of the situation that is closer to reality. If we trace back the history of Artificial Intelligence, some may be surprised that this field stems from logical reasoning and knowledge-based approaches like the expert system until the industry busts and the “AI winter” in the 1990s. At that time, black-box models are not yet the mainstream. It is the rise of statistical approaches after that leads to the situation now. If we look back, we can see that the success of ML now is a product of a combination of huge amount of data, much more

powerful computing power, and more advanced AI algorithms. This wave is much more welcoming for the black-box systems than before. Building off these exciting new techniques, more applications are developed: autonomous driving, smart city, medical assistance, etc. Just when the future seems brighter than ever, things again turned in the other direction. As people employ these black-box models as increasingly important and safety-critical components, we also witness both benefits and harms from such technologies. A prime example is the accident of Uber's self-driving car, which resulted in the death of Elain Herzberg, the first pedestrian to be killed by an autonomous vehicle. Regardless of what caused the accident from a technological perspective, this event started a heated social discussion about whether we have let the technology loose in negligence of safety and transparency issues. Now, more and more efforts from both the academia and the industry are spent on enhancing transparency to strengthen the safety of such systems. This is the large context that motivates me to conduct this STS research project, and my specific research questions are:

- how to frame these two sides of thinking?
- how does this dichotomy between efficiency and transparency come into being?
- where does the conflicts between them emerge, and if they can be resolved?

To attempt answering these questions, I believe understanding the constructivist side of the story is as important as the mathematics and algorithms because it reveals the full picture and can potentially clarify what is needed to improve the technologies in the future.

A fitting STS theory for this project is the theory of technological momentum proposed by the historian of technology Thomas P. Hughes. This is a theory about the relationship between technology and society over time. In its essence, Hughes's theory is a synthesis of two

models, technological determinism and social determinism, for how technology and society interact. Technological determinism claims that society itself is modified by the introduction of a new technology in an irreversible and irreparable way, and technology, under this model, self-propagates as well—there is no turning back once the adoption has taken place, and the very existence of the technology means that it will continue to exist in the future. On the other hand, social determinism claims that society itself controls how a technology is used and developed. Technological momentum tries to unify the two models by adding time. In Hughes's theory, when a technology is young, deliberate control over its use and scope is possible and enacted by society. However, as a technology matures, and becomes increasingly enmeshed in the society where it was created, its own deterministic force takes hold, achieving technological momentum in the process. According to Hughes this inertia, which is particularly the case for large technological systems with their technological and social components, makes them difficult to influence and steer as they start to go more on their own way, assuming deterministic traits in the process (Vermaas et al., 2011). In other words, Hughes's says that the relationship between technology and society always starts with a social determinism model, but evolves into a form of technological determinism over time and as its use becomes more prevalent and important. Does this theory satisfactorily explain the development of AI and ML technologies? Or do we need more tools in order to properly address the research questions?

To define the scope of the project, I try to identify the relevant stakeholders in this problem, including both human and non-human actors. The human actors are grouped as followed:

- the public: the users, consumers of AI technologies; the non-users, people affected by the technologies

- the industry: the AI company stakeholders, businessman employing these technologies
- the academia: various scientific groups with different epistemologies, e.g. data-driven/rule-based, empiricism/rationalism, etc.

Interestingly, I have observed a divergence of attitudes toward the use of black-box models among these three groups of stakeholders. In this project, I will focus more on the human actors, and an interesting part of the investigation will be devoted to the analysis of how the interaction between human actors influence the choice of technology. The non-human actors are the list of philosophies, mathematical, algorithmic techniques, which are more important to the academia, and the technical systems or artifacts, which are more important to the industry and the public.

D. Conclusion

The end objective of my STS research is to build an argument for the explanation of why, where and how the conflicts between two AI directions, efficiency and transparency, emerge. To concertize the argument, I will use two case studies to illustrate the emergence of conflicts from two different periods of time. Throughout the project, my research method will be mainly historical, which relies on the literature reviews of different actors listed above. I have outlined the tentative timeline for each section of the final paper in Table 1.

My goal for this STS research project is to try to uncover the forces at play in the "efficiency vs transparency" of black-box systems, frame them in appropriate terms, study how they interact and evolve to this day, and finally predict the potential impact this dynamics have on the technologies and the society.

References

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. The MIT Press.

Vermaas, P., Kroes, P., Franssen, M., van de Poel, I., & Houkes, W. (2011). *A Philosophy of Technology: From Technical Artefacts to Sociotechnical Systems*. Morgan & Claypool Publishers.

Thrun, S., Burgard, W., & Fox, D. (2005). Probabilistic Robotics (In-telligent Robotics and Autonomous Agents). In *Algorithmica (New York)*. The MIT Press.

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, 279–288. https://doi.org/10.1145/3287560.3287574

Lacerda, B., Faruq, F., Parker, D., & Hawes, N. (2019). Probabilistic planning with formal performance guarantees for mobile service robots. *The International Journal of Robotics Research*, *38*(9), 1098–1123.

Turek, M. (2016). Explainable Artificial Intelligence (XAI). Retrieved from Defense Advanced Research Projects Agency (DARPA) website: https://www.darpa.mil/program/explainable-artificial-intelligence

Goebel, R., Chander, A., Holzinger, K., Lecue, F., Stumpf, S., Kieseberg, P., … Lecue, F. (2018). Explainable AI : the new 42 ? To cite this version : HAL Id : hal-01934928. *2nd International Cross-Domain Conference for Machine Learning and Knowl- Edge Extraction (CD-MAKE)*.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007