Towards Trustworthy Autonomous Systems via Conversations and Explanations

Balint Gyevnar

School of Informatics
University of Edinburgh
balint.gyevnar@ed.ac.uk

Abstract

Autonomous systems fulfil an increasingly important role in our societies, however, AI-powered systems have seen less success over the years, as they are expected to tackle a range of social, legal, or technological challenges and modern neural network-based AI systems cannot yet provide guarantees to many of these challenges. Particularly important is that these systems are black box decision makers, eroding human oversight, contestation, and agency. To address this particular concern, my thesis focuses on integrating social explainable AI with cognitive methods and natural language processing to shed light on the internal processes of autonomous systems in a way accessible to lay users. I propose a causal explanation generation model for decision-making called CEMA based on counterfactual simulations in multi-agent systems. I also plan to integrate CEMA with a broader natural language processing pipeline to support targeted and personalised explanations that address people's cognitive biases. I hope that my research will have a positive impact on the public acceptance of autonomous agents by building towards more trustworthy AI.

Introduction

From routing network packets to sorting warehouses and driving cars, autonomous systems are an important presence in today's society. However, many of these systems operate in safety- and privacy-critical environments. Consequently, modern Artificial Intelligence (AI) based on deep learning or reinforcement learning has had less success in these fields, as people seem reluctant to adopt these systems into their lives.

This lack of trust is not surprising as numerous issues need to be addressed urgently. We must take into account the complex socio-technical interactions that autonomous systems are expected to tackle. Only by building systems that address these interactions can we hope to build *trustworthy AI*. Our focus should include issues such as diversity, fairness, societal and environmental well-being, technical robustness, safety, privacy, and human decision-making agency. Rather than merely improving an arbitrary measure of trust, the purpose of trustworthy autonomous systems should, therefore, be to enshrine and protect basic human rights and enable sustainable innovation (Gyevnar, Ferguson, and Schafer 2023). While neural AI systems offer impressive performance, they are also black boxes. Reliance on such systems hinders people's ability to contest decisions and affects their decision-making autonomy. Therefore, we should strive to increase the transparency of these systems to restore human agency and enable contestability.

The field of explainable AI (XAI) has long been trying to shed light on the inner workings of AI systems, however, the need for trustworthy AI has promoted a shift towards a more social approach (Miller 2019). Traditional XAI is useful so long as the "explanations" – usually some relative ordering of features, saliency maps, or attention weights – are observed by experts. In contrast, social XAI focuses on intelligible explanations that reveal the causes behind the decisions of an autonomous system. Further, these explanations are tailored to take into account human cognitive biases and appeal to the social nature of humans through conversations

My research is in this field of social XAI. The goal is to build a framework that can deliver easy-to-understand natural language explanations to people's queries about any autonomous system making sequential decisions in a multi-agent environment. The explanations are to be delivered in terms of causes behind the decisions of the agent as part of a dialogue system that can keep track of and update an internal model of people's knowledge about the autonomous system, thereby targeting the right cognitive requirements of users.

Causal Selection for Autonomous Driving

Explanations by "opening the black box" of large neural models, that is, using knowledge about the intrinsic properties of the system, are often not feasible (Wachter, Mittelstadt, and Russell 2017). One way to address this is through contrastive explanations which can be generated by varying only the inputs to the system. Contrastive explanations are also causally grounded as they highlight counterfactual features in a system that affect the outcome. In addition, they are also better aligned with how humans explain causal relationships (Miller 2019). A large body of literature focuses on contrastive explanations of machine learning (Stepin et al. 2021) and some methods have been proposed for deterministic single-agent environments and well-defined planning domains (Chakraborti, Sreedharan, and Kambhampati 2021). However, sequential decision-making in dynamic and coupled multi-agent systems has received less attention.

My initial work started with mapping the decisions of an autonomous vehicle (AV) to a Bayesian network to extract the causes behind the decisions of the AV (Gyevnar et al. 2022). However, the generated explanations of this system proved to be too high-level and so I needed to find a more expressive approach. After several failed attempts, I came across the work of Quillien and Lucas (2023) which provides an empirically validated account of how humans themselves may select causes for their explanations It is called the Counterfactual Effect Size Model and its assumptions are few and intuitive. First, it assumes that people sample from a cognitive distribution across counterfactual worlds that are grounded in the actual observations of the world. Second, it assumes that people calculate causal effect size by correlating features with outcomes across counterfactuals. When one outcome is present if and only if one feature is present, then that feature is assigned a large causal effect.

Based on this model, we proposed a system called Causal Explanations in Multi-Agent systems (CEMA) (Gyevnar et al. 2023). Unlike my previous work, CEMA is applicable to explain the decisions of an *ego agent* in any multi-agent system. It relies on a probabilistic model that can predict the subsequent states of the environment conditioned on previous states. In other words, instead of *a priori* assuming a fixed causal structure, CEMA uses simulations to extract causes for which auto-regressive models or probabilistic policies are widely available. CEMA can also generate low- and high-level explanations, as we do not assume a specific structure or abstraction over the states of the environment.

To address further requirements of social XAI, we designed CEMA with user interaction at its core. Users pose queries about the actions of the ego to which CEMA delivers selected and relevant causal explanations in three main steps. First, the current state of the world is rolled back to some past time, erasing the queried actions of the ego agent. From this past time, the probabilistic model is used to sample a set of counterfactual worlds, providing information about with which features of the world the queried actions of the ego agent co-occur. Finally, a measure of correlation is calculated between features of the world and the queried actions of the ego vehicle, ranking them by their counterfactual causal effect size. This causal selection process is shown in Figure 1.

We evaluate CEMA for motion planning in autonomous driving using four scenarios with complex interactions. We show that it identifies correct and relevant causes in all scenarios even when a large number of causally irrelevant agents are present. We also perform a user study (N=200) using CEMA's explanations and show that participants rank them at least as high as baseline explanations elicited from other human participants. The user study also indicates that CEMA's explanations positively affect people's level of trust.

Future Work

Initial results with CEMA are reassuring but there is much left to be done. First, the current method is focused on causal selection, but the integration with modern conversational agents and the tracking of users' mental models is essential for social XAI. For this, my following step is to integrate CEMA with language models to seamlessly parse human queries and convert causal information to fluent natural language sentences while conditioning on updated knowledge about the users' mental models. I will also evaluate CEMA in other domains not just autonomous driving to conclusively show that it is indeed applicable in any multi-agent system.

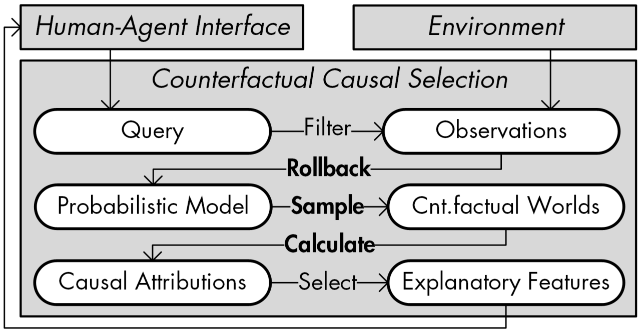

Figure 1: The structure of our causal explanation framework, Causal Explanations in Multi-Agent systems (CEMA)

References

Chakraborti, T.; Sreedharan, S.; and Kambhampati, S. 2021. The Emerging Landscape of Explainable AI Planning and Decision Making. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, 4803–4811. Yokohama, Yokohama, Japan. ISBN 978-0-9992411-6-5.

Gyevnar, B.; Ferguson, N.; and Schafer, B. 2023. Bridging The Transparency Gap: What Can Explainable AI Learn From The AI Act? In *Proceedings of ECAI 2023, the 26th European Conference on Artificial Intelligence*, Frontiers in Artificial Intelligence and Applications, 964 – 971. IOS Press.

Gyevnar, B.; Tamborski, M.; Wang, C.; Lucas, C. G.; Cohen, S. B.; and Albrecht, S. V. 2022. A Human-Centric Method for Generating Causal Explanations in Natural Language for Autonomous Vehicle Motion Planning. In *IJCAI 2022 Workshop on Artificial Intelligence for Autonomous Driving*.

Gyevnar, B.; Wang, C.; Lucas, C. G.; Cohen, S. B.; and Albrecht, S. V. 2023. Causal Explanations for Sequential Decision-Making in Multi-Agent Systems. In *IJCAI 2023 Workshop on Explainable Artificial Intelligence*, arXiv:2302.10809.

Miller, T. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, 267: 1–38.

Quillien, T.; and Lucas, C. G. 2023. Counterfactuals and the Logic of Causal Selection. *Psychological Review*, Advance Online Publication.

Stepin, I.; Alonso, J. M.; Catala, A.; and Pereira-Fariña, M. 2021. A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence. *IEEE Access*, 9: 11974–12001.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology (Harvard JOLT)*, 31(2): 841–888.