Building Trustworthy Human-Centric Autonomous Systems Via Explanations

Doctoral Consortium

Balint Gyevnar
University of Edinburgh
Edinburgh, United Kingdom
balint.gyevnar@ed.ac.uk

ABSTRACT

Autonomous systems suffer from people's mistrust, as these systems rely on highly accurate yet inscrutable black box methods that are not amenable to safety guarantees nor common sense understanding. As a result, we see the erosion of accountability, human oversight, and contestation. In an attempt to build transparency, I advocate the use of model-specific, interactive, intelligible, and causally-grounded explanations for autonomous systems that take the human factor into account. I proposed a simulation-based conversational and causal framework for explaining sequential decision-making. The method, which is called CEMA, satisfies the previous four criteria without sacrificing the performance of complex models. I verified the benefits of CEMA via extensive quantitative and qualitative evaluation involving a large user study and autonomous driving. However, future work remains. To build a trustworthy autonomous system, CEMA needs to provide explanations that accurately calibrate people's trust according to the capabilities of the system. Towards this end, I hope to exploit prior knowledge in large language models to extend CEMA into a trust calibration system that uses conversations and explanations to adjust people's trust appropriately.

KEYWORDS

trustworthy autonomous systems; explainable AI; conversational agents; causal explanations; trust calibration; autonomous driving

ACM Reference Format:
Balint Gyevnar. 2024. Building Trustworthy Human-Centric Autonomous Systems Via Explanations: Doctoral Consortium. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

Artificial intelligence-based autonomous systems have huge potential to solve complex tasks and many of these systems physically interact with users in safety- or privacy-critical environments. However, modern deep learning or reinforcement learning-based methods have had less success in these areas, as they often lack the necessary robustness and safety guarantees necessary to develop an appropriate level of trust in users. Unfortunately, this mistrust is symptomatic of more significant issues that underlie the deployment of complex AI systems. While these methods offer impressive performance, to most users they are black boxes that are inscrutable. Reliance on these systems hinders users' ability to understand and contest decisions, which limits their decision-making agency.

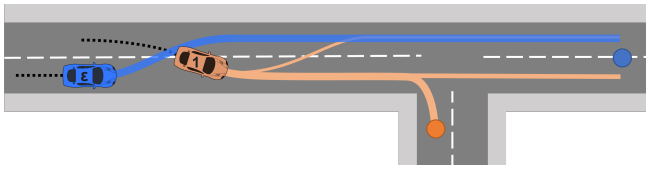

Figure 1: The autonomous vehicle is heading to the blue goal. It changed lanes after the other vehicle cut in front of it and slowed down. A passenger asks: *Why did you change lanes?* It was faster because the vehicle in front was slower than us. *Was it safe to change lanes?* Yes, there was no one in the next lane. *What if you had missed a vehicle?* If we were uncertain about the environment, we would remain in our lane.

As a consequence, it has become important to build methods that explicitly address this need for *transparency*, and so explainable AI (XAI) has gained prominence. However, traditional post-hoc interpretability methods of XAI are only useful so long as the "explanations" – usually some relative ordering of features, saliency maps, or attention weights – are observed by domain or AI experts. In addition, autonomous systems deployed in safety- and privacy-critical real-world environments need to be able to guarantee model-specific safety and calibrate people's trust accurately and according to the system's capabilities. This requires a very different approach from traditional post-hoc model-agnostic XAI methods. One potential way forward is to focus on a more *social XAI* [5] with intelligible explanations that reveal the causes behind the decisions of autonomous systems. These explanations are tailored to take into account people's cognitive biases while appealing to the social nature of humans through conversations.

However, achieving explainability does not commute with achieving transparency. The former is a means necessary to achieve the latter whose goal is to restore people's decision-making agency [1]. There is often a tacit understanding in XAI, that explainability equals transparency, but the unfortunate nature of this standpoint is reflected in the futility of trying to establish post-hoc model-agnostic methods as "trust building" tools [7, 9]. However, by remembering that XAI is just a means to achieve transparency – as much as documentation, auditing, or standardisation are means towards it – we can still make great of use both traditional and

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

social XAI methods as long as the design choices are justified in terms of domain- and stakeholder-specific requirements.

It is against this backdrop, that I situate my research in the field of social XAI. My goal is to build a framework – illustrated in Figure 1 for the domain of autonomous driving (AD) – that delivers easy-to-understand natural language explanations to people's queries about any autonomous system that makes sequential decisions in a multi-agent environment. The explanations are delivered in terms of the causes behind the decisions of the agent as part of a conversation that keeps track of and updates an internal model of people's knowledge about the autonomous system, thereby aiming to accurately calibrate people's trust levels.

2 TRUSTWORTHY EXPLANATIONS FOR MULTI-AGENT SYSTEMS

There are three main tasks to calibrating trust with explanations from the perspective of social XAI: (1) identifying pertinent causal factors behind the agent's decisions; (2) delivering intelligible explanations that address the user's concerns; (3) evaluating the effects of explanations on trust with the actual stakeholders.

For the first task, note that "opening the black box" of deep learning, that is, using knowledge about the intrinsic properties of the system, are often not feasible [10]. Instead, we can rely on the theory of counterfactual causation to extract causes in the context of hypothetical scenarios, thus shedding light on the alternative behaviours of our systems [3, 4]. However, doing this for complex and dynamically evolving multi-agent systems is a great challenge.

Following experimentation with unsuccessful designs, I found the work of Quillien and Lucas [8] which provides an empirically validated account of how humans themselves may select causes for their explanations. This is called the Counterfactual Effect Size Model (CESM) and it is based on two assumptions. People sample from a cognitive distribution across counterfactual worlds grounded in the observations of the factual world, and they calculate causal effect size by correlating features and outcomes across counterfactuals. When an outcome is present if and only if one feature is present, then that feature is assigned a large causal effect.

Based on this, I proposed the CEMA system which stands for Causal Explanations in Multi-Agent systems [2] and is applicable to explain the decisions of any *ego agent* in a multi-agent system. It uses simulations from a probabilistic model of the subsequent states of the environment to create counterfactuals that are used within the CESM framework to determine the effect size of both intrinsic teleological and extrinsic mechanistic causes. This allows CEMA to use existing decision-making models, stochastic policies, etc. for generating local causal explanations without *a priori* assuming something explicitly about the causal structure in the world.

In line with the requirements of social XAI, I also designed CEMA to support model-specific and intelligible conversations. Users pose *queries* about the actions of the ego agent to which CEMA delivers responses in terms of domain-specific explanations that may be constructed from both low- and high-level features of the environment. This is done in three main steps shown in Figure 2. First, the current state of the world is rolled back to the past, erasing the queried actions of the ego agent. From then, CEMA simulates a set of counterfactual worlds. This provides information about the features of the world with which the queried actions of the ego co-occur. Finally, a measure of correlation is calculated between features of the world and the queried actions of the ego vehicle. This ranks features by their counterfactual causal effect size.

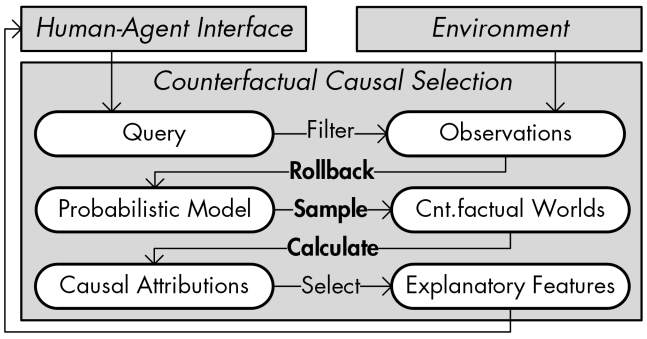

Figure 2: The structure of our causal explanation framework, Causal Explanations in Multi-Agent systems (CEMA)

I evaluated CEMA using the task of motion planning for autonomous driving with four scenarios with coupled agent interactions, showing that CEMA identifies correct and relevant causes in all of the scenarios even when a large number of irrelevant agents are present. I also performed a user study (N=200) using CEMA's explanations that showed that participants ranked them for correctness and relevancy at least as high as baseline explanations elicited from other human participants.

3 GOING FORWARD

I have mostly only addressed the first of three tasks set out at the start of the previous section. While CEMA is formulated as a conversational process, my focus was largely on just the causal selection aspect. I could not sufficiently address the goal of trust calibration or language processing. One possible direction forward is to leverage the latent prior knowledge stored in large language models (LLM) to predict when and about what users require explanations.

Users should be able to query and give feedback directly to the system which can then track whether they under- or over-trust it. The goal of trust calibration is not to make people trust a system more but to make sure that they are aware of the system's capabilities and know when and when not to trust it. Integration with LLMs would not only enable this process, but it would also provide natural language processing capabilities that are necessary to perform this calibration. One way to do this in practice might be to provide a structured verbalisation of the scene to the LLM at the start, and as the user queries CEMA, we gradually update this context with the extracted causes. We may then prompt the user for a self-assessment of their trust in the system. In conjunction with our existing contextual information stored in the LLM, this could then further refine whether and what sort of explanations are still necessary for the user. This could also form the basis of an interactive evaluation where users can experience the system and which should measure the trustworthiness of the system in ways that avoids the pitfalls of narrowly focused "trust-building" [6].

REFERENCES

[1] Balint Gyevnar, Nick Ferguson, and Burkhard Schafer. 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)*. IOS Press, 964 – 971. https://doi.org/10.3233/FAIA230367

[2] Balint Gyevnar, Cheng Wang, Christopher G. Lucas, Shay B. Cohen, and Stefano V. Albrecht. 2024. Causal Explanations for Sequential Decision-Making in Multi-Agent Systems. In *23rd International Conference on Autonomous Agents and Multi-Agent Systems*. IFAAMAS.

[3] Denis J. Hilton. 1988. Logic and Causal Attribution. In *Contemporary Science and Natural Explanation: Commonsense Conceptions of Causality*. New York University Press, New York, NY, US, 33–65.

[4] David Lewis. 1973. Causation. *Journal of Philosophy* 70, 17 (1973), 556–567. https://doi.org/10.2307/2025310

[5] Tim Miller. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence* 267 (Feb. 2019), 1–38.

[6] Tim Miller. 2022. Are We Measuring Trust Correctly in Explainability, Interpretability, and Transparency Research? arXiv:2209.00651 [cs]

[7] Tim Miller. 2023. Explainable AI Is Dead, Long Live Explainable AI! Hypothesis-driven Decision Support Using Evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 333–342. https://doi.org/10.1145/3593013.3594001

[8] Tadeg Quillien and Christopher G. Lucas. 2023. Counterfactuals and the Logic of Causal Selection. *Psychological Review* (2023), Advance Online Publication.

[9] Cynthia Rudin. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence* 1, 5 (May 2019), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[10] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology (Harvard JOLT)* 31, 2 (2017), 841–888.