Accountability Layers: Explaining Complex System Failures by Parts

Leilani H. Gilpin

UC Santa Cruz
lgilpin@ucsc.edu

Abstract

With the rise of AI used for critical decision-making, many important predictions are made by complex and opaque AI algorithms. The aim of eXplainable Artificial Intelligence (XAI) is to make these opaque decision-making algorithms more transparent and trustworthy. This is often done by constructing an "explainable model" for a single modality or subsystem. However, this approach fails for complex systems that are made out of multiple parts. In this paper, I discuss how to explain complex system failures. I represent a complex machine as a hierarchical model of introspective subsystems working together towards a common goal. The subsystems communicate in a common symbolic language. This work creates a set of explanatory *accountability layers* for trustworthy AI.

As AI systems take control of decisions previously entrusted to humans, e.g., navigation, or driving, society needs trustworthy assessments. One way to assess these AI systems before deployment is to use an explanation. But these explanations local to a specific subsystem, component or part. When using explanations in autonomous agents, composed of multiple modalities, components, and parts, we need (1) a hierarchical model of (2) multiple explanatory subsystems.

Accountability Layers represent an agent as a hierarchical model of introspective subsystems working together towards a common goal. This idea builds upon my previous work of using *explanations in two distinct ways*: (i) system-wide explanations provided to an end-user for analysis (Gilpin 2020; Gilpin, Penubarthi, and Kagal 2021), and (ii) internal explanations used among subsystems to defend their actions (Gilpin 2018; Gilpin, Macbeth, and Florentine 2018) and make more robust higher-level decisions. Therefore, I differentiate between (i) an internal subsystem explanation (or internal explanation), defined using symbolic reasons and dependencies for a specific, local subsystem's behavior, and (ii) a system-wide narrative explanation (or system explanation), composed of a (mostly causal) chain of reasoning generated from the underlying subsystems. Since the underlying reasons and dependencies are symbolic, they can be translated into a human-understandable explanation at various degrees of detail from low-level (e.g., anomaly detection) to high-level (e.g., failure diagnosis). Previously, these ideas were applied to autonomous vehicles (Gilpin 2020; Gilpin, Penubarthi, and Kagal 2021), a technology in which safety is of paramount importance for the protection of human lives. Ongoing work extends this work to (1) be adaptable to multiple types of embodied agents (including but not limited to self driving cars) and (2) to include neuro-symbolic reasoning. In summary, my research work is motivated by the following question:

Can we develop self-explaining architectures that can help anticipate failures instead of providing justifications post hoc?

When explanations are wrong, there is no way to correct them. This is a necessity for safety-critical systems like autonomous vehicles where opaque vision systems have led to passenger injuries and pedestrian deaths without an explanation of how and why. In the human analog, when someone provides an explanation that is incorrect, inappropriate, or unrelated, the explanation can be corrected, changed or even retracted. My motivation is to represent machine-generated explanations as layers of information. There may be additional layers added to for (1) understanding and interpretation, (2) interaction and annotation or (3) quantification and confidence of the underlying explanation. The societal need for explanations that we can trust motivates the intellectual merit of this research. Accountability layers will allow users to ascertain their level of trust in the method and appropriately weigh their contribution in critical decision making.

Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

References

Gilpin, L. H. 2018. Reasonableness Monitors. In *The Twenty-Third AAAI/SIGAI Doctoral Consortium at AAAI-18*. New Orleans, LA: AAAI Press.

Gilpin, L. H. 2020. *Anomaly detection through explanations*. Ph.D. thesis, Massachusetts Institute of Technology.

Gilpin, L. H.; Macbeth, J. C.; and Florentine, E. 2018. Monitoring Scene Understanders with Conceptual Primitive Decomposition and Commonsense Knowledge. *Advances in Cognitive Systems*, 6.

Gilpin, L. H.; Penubarthi, V.; and Kagal, L. 2021. Explaining Multimodal Errors in Autonomous Vehicles. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, 1–10. IEEE.