Formal Explainability for Artificial Intelligence in Dynamic Environments

Jime Cuartas Granada

Supervisors:
Alexey Ignatiev
Peter J. Stuckey
Julian Gutierrez

A Progress Review report at
Monash University in 2026
School of Information Technology

Abstract

In dynamic environments, the goal of Artificial Intelligence (AI) is to build intelligent agents capable of addressing sequential decision-making settings. Reinforcement Learning (RL) is a branch of Machine Learning that addresses sequential decision-making by agents to perform tasks. In this context, there are two important challenges for humans to understand decisions made by agents: (1) the sequential decisions are connected, and (2) the agents may use opaque black-box models (e.g., neural networks) for each decision.

Despite the success of RL in sequential decision-making, the lack of transparency in understanding their decisions can make the agents hard to validate. To address the need for transparency, there are efforts to develop Explainable Artificial Intelligence (XAI) and its subfield, Explainable Reinforcement Learning. XAI is a set of methods designed to make AI models easier to comprehend. Despite the importance of Explainable Reinforcement Learning in developing trustworthy intelligent agents, there are gaps in current research to make sequential decision-making explainable.

This project proposes to explain sequential decision-making using formal reasoning. To achieve this goal, the proposal focuses on (1) Formal Explainability for Finite Automata, to address sequential actions in deterministic environments, and (2) Formal Explainability for Reinforcement Learning, where the agent's behaviour is non-interpretable.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The deplyoyment of Artificial Intelligence (AI) algorithms has necessitated the need for eXplainability AI (XAI) methods in order to ensure transparency, trust, and accountability. While much of the field has focused on heuristic explanations for opaque models, there is an interest in formal approaches that provide rigorous guarantees about the explanations generated [1, 2].

A fundamental challenge in dynamic environments is explaining sequential decision-making, To address this, we model these processes using Automata, which provide a symbolic and tractable representation of sequential decision functions. This approach allows us to generate formal explanations, why a specific sequence of actions leads to a particular outcome. Automata are widely used in software verification [3], design of communication protocols [4],and sintax parsing in compiler [5]. When a computational model, such as a Finite Automaton (FA) or a Pushdown Automaton (PDA), accepts or rejects an input string, the reasoning behind that decision can be non-trivial. Understanding why a specific input was accepted or rejected is crucial for debugging, and refinement purposes.

This research project investigates the formalization of explanations for sequential decision-making. Having addressed an approach to deliver formal explanations for Finite Automata (FA) in the first stage of this research, and submitting it to a ICALP 2026. We now move to address explanations for Context-Free Languages (CFG) using Pushdown Automata (PDA).

1.1 Problem Statement

While standard XAI focuses on feature attribution in classifiers, the "features" in formal languages are sequential and structural. Since the confirmation report, the research scope has been refined to address three primary gaps:

1.2 Problem Statement

While standard XAI focuses on feature attribution in classifiers, the "features" in formal languages are sequential and structural. Since the confirmation report, the research scope has been refined to address three primary gaps:

- **Research Problem 1 (Completed): Explaining Finite Automata.** Finite Automata are often assumed to be interpretable. However, large FA are cognitively inaccessible to humans. We have developed a framework to compute formal explanations for the acceptance and rejection of inputs in FA, providing a rigorous foundation for automaton-based explainability.

- **Research Problem 2: Explaining Pushdown Automata (PDA).** Context-free languages, recognized by PDAs, introduce a stack-based memory that allows to represent makes explanations more complex. A single character's "badness" may depend on a token seen much earlier in the stream. The second problem addresses the generation of Minimal Contrastive Explanations (CXPs) the minimal sets of modifications required to turn a rejected word into an accepted one.

  There is a lack of quantitative metrics that assign a "degree of responsibility" to specific indices in a rejected string. The third problem focuses on the development of the Features Attribution Score (RAS), using constrained optimization (Non-Negative Least Squares) to provide a probabilistic ranking of which tokens most significantly contribute to a structural rejection.

- **Research Problem 3: Explaining Markov Decision Processes.** How can the Feature Attribution Score be extended to explain failure states in Reinforcement Learning policies modeled as MDPs? This problem explores the adaptation of RAS to sequential decision-making, where actions influence future states and rewards.

1.3 Responsibility Attribution for Token Substitutions

The linear system treats the probability of a CXP as a shared resource among its constituent indices. For each index i, the "Responsibility Score" S_i is:

$$S_i = \sum_{j:i \in cxp_j} \frac{P(cxp_j)}{|cxp_j|}$$

Where $|cxp_j|$ is the size of the explanation (minimal set). This accounts for the fact that if an explanation requires changing 5 tokens, the credit for that fix is diluted across those 5 letters

1.4 Badness Attribution

Para justificar el uso de la multiplicación en tu modelo de Rejection Attribution Score (RAS) y su relación con la "maldad" (v_i) en el contexto de tu doctorado, debes fundamentarlo en tres pilares: la lógica booleana de los autómatas, la teoría de probabilidad de las gramáticas (PCFG) y la tratabilidad matemática. Aquí tienes la justificación técnica y académica:1. La lógica del "Y" (Intersección de Eventos)En la teoría de lenguajes formales, para que una palabra sea aceptada por un autómata (o generada por una gramática), todas las restricciones estructurales deben cumplirse simultáneamente. Es una condición de tipo AND.Si definimos $(1 - v_i)$ como la probabilidad de que el índice i sea "correcto" o "válido" según las reglas de la gramática, la aceptación de la palabra completa es el evento donde el índice 1 es válido Y el índice 2 es válido Y así sucesivamente.En teoría de probabilidad, la probabilidad de la intersección de eventos independientes es el producto de sus probabilidades individuales: $P(\text{Aceptación}) = P(\text{validez}_1) \cdot P(\text{validez}_2) \ldots$ Por lo tanto, $P = \prod(1 - v_i)$ es la representación natural de una estructura donde el fallo de un solo componente puede comprometer la aceptación total.

2. Justificación mediante el modelo de "Canal con Ruido"Para tu tesis, puedes presentar la palabra rechazada como una versión "ruidosa" o "corrupta" de una palabra válida.Justificación: Cada v_i representa la probabilidad de que haya ocurrido un error de transformación en el índice i. Bajo esta premisa, la probabilidad de que una palabra reparada (una explicación contrastiva como $[1, 3]$) sea la "verdadera" intención del

usuario depende de que los demás índices se mantengan correctos. La multiplicación cuantifica la verosimilitud de esa reparación estructural.

Appendix A

An Appendix

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Vivamus at pulvinar nisi. Phasellus hendrerit, diam placerat interdum iaculis, mauris justo cursus risus, in viverra purus eros at ligula. Ut metus justo, consequat a tristique posuere, laoreet nec nibh. Etiam et scelerisque mauris. Phasellus vel massa magna. Ut non neque id tortor pharetra bibendum vitae sit amet nisi. Duis nec quam quam, sed euismod justo. Pellentesque eu tellus vitae ante tempus malesuada. Nunc accumsan, quam in congue consequat, lectus lectus dapibus erat, id aliquet urna neque at massa. Nulla facilisi. Morbi ullamcorper eleifend posuere. Donec libero leo, faucibus nec bibendum at, mattis et urna. Proin consectetur, nunc ut imperdiet lobortis, magna neque tincidunt lectus, id iaculis nisi justo id nibh. Pellentesque vel sem in erat vulputate faucibus molestie ut lorem.

Quisque tristique urna in lorem laoreet at laoreet quam congue. Donec dolor turpis, blandit non imperdiet aliquet, blandit et felis. In lorem nisi, pretium sit amet vestibulum sed, tempus et sem. Proin non ante turpis. Nulla imperdiet fringilla convallis. Vivamus vel bibendum nisl. Pellentesque justo lectus, molestie vel luctus sed, lobortis in libero. Nulla facilisi. Aliquam erat volutpat. Suspendisse vitae nunc nunc. Sed aliquet est suscipit sapien rhoncus non adipiscing nibh consequat. Aliquam metus urna, faucibus eu vulputate non, luctus eu justo.

Donec urna leo, vulputate vitae porta eu, vehicula blandit libero. Phasellus eget massa et leo condimentum mollis. Nullam molestie, justo at pellentesque vulputate, sapien velit ornare diam, nec gravida lacus augue non diam. Integer mattis lacus id libero ultrices sit amet mollis neque molestie. Integer ut leo eget mi volutpat congue. Vivamus sodales, turpis id venenatis placerat, tellus purus adipiscing magna, eu aliquam nibh dolor id nibh. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Sed cursus convallis quam nec vehicula. Sed vulputate neque eget odio fringilla ac sodales urna feugiat.

Phasellus nisi quam, volutpat non ullamcorper eget, congue fringilla leo. Cras et erat et nibh placerat commodo id ornare est. Nulla facilisi. Aenean pulvinar scelerisque eros eget interdum. Nunc pulvinar magna ut felis varius in hendrerit dolor accumsan. Nunc pellentesque magna quis magna bibendum non laoreet erat tincidunt. Nulla facilisi.

Duis eget massa sem, gravida interdum ipsum. Nulla nunc nisl, hendrerit sit amet commodo vel, varius id tellus. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nunc ac dolor est. Suspendisse ultrices tincidunt metus eget accumsan. Nullam facilisis, justo vitae convallis sollicitudin, eros augue malesuada metus, nec sagittis diam nibh ut sapien. Duis blandit lectus vitae lorem aliquam nec euismod nisi volutpat. Vestibulum ornare dictum tortor, at faucibus justo tempor non. Nulla facilisi. Cras non massa nunc, eget euismod purus. Nunc metus ipsum, euismod a consectetur vel, hendrerit nec nunc.

Bibliography

[1] João Marques-Silva. Logic-based explainability in machine learning. In Leopoldo E. Bertossi and Guohui Xiao, editors, *Reasoning Web. Causality, Explanations and Declarative Knowledge - 18th International Summer School 2022, Berlin, Germany, September 27-30, 2022, Tutorial Lectures*, volume 13759 of *Lecture Notes in Computer Science*, pages 24–104. Springer, 2022. doi: 10.1007/978-3-031-31414-8_2. URL https://doi.org/10.1007/978-3-031-31414-8_2.

[2] Adnan Darwiche. Logic for explainable AI. In *LICS*, pages 1–11. IEEE, 2023.

[3] Christel Baier and Joost-Pieter Katoen. *Principles of model checking*. MIT Press, 2008. ISBN 978-0-262-02649-9.

[4] Gerard J. Holzmann. The model checker SPIN. *IEEE Trans. Software Eng.*, 23 (5):279–295, 1997. doi: 10.1109/32.588521. URL https://doi.org/10.1109/32.588521.

[5] Alfred V. Aho, Ravi Sethi, and Jeffrey D. Ullman. *Compilers: Principles, Techniques, and Tools*. Addison-Wesley series in computer science / World student series edition. Addison-Wesley, 1986. ISBN 0-201-10088-6. URL https://www.worldcat.org/oclc/12285707.