Conversational Agents and eXplainable Artificial Intelligence

Antoine Blancy, Thomas Dagier

AUI 2024

Introduction - Context

Artificial Intelligence keeps getting better

- Black boxes =/= trust or confidence

Chatbots are getting more popular

- AI + Chatbot = profit?

Introduction - Explainable AI, Why?

Importance in high-impact scenarios

- Healthcare, finance, law

Understanding = trust!

Introduction - Definitions

Local explanations:

- Model’s behavior for individual predictions

Global explanations:

- Model’s behavior across the entire dataset
  - Patterns, biases, etc.

Introcution - Definitions

Explainable AI = 3 main components

Introduction - Definitions

Interpretability

- Definition:
  - Understand the inner workings of a model
- Techniques:
  - Feature attribution
  - Attention mechanisms

Introduction - Definitions

Understandability

- Definition:
  - Understanding the decision-making process for a given input
- Methods:
  - Natural language explanations

Introduction - Definitions

Accountability

- Definition:
  - Ability to trace actions and decisions made by a model
- Techniques:
  - Logging decision pathways

Introduction - Conversational Agent

Can be anything the user interacts with

In our case: chatbots

Introduction - Project Goals

Create a chatbot with enhanced explainability

For more trust and transparency

Conversational Agent Design

Trust factors:

- **Human likeness**
- **Visual design**

Tools for Explainability

AI Explainability 360 (AIX360)

- Open-source library
- Algorithms for interpretability and explainability
- Supports text, images, time series
- Local post-hoc explanations & global explanations

ELI5 (Explain Like I’m 5)

- Python package
- Debug and explain machine learning classifiers
- Supports various ML frameworks (scikit-learn, keras, etc.)

InterpretML

- Open-source toolkit
- Tools for blackbox and glassbox models
- Interactive visualizations to help explain decisions
- Particular focus on NLP models

Dalex2

- Open-source R library
- Similar to InterpretML
- Tools for understanding model behavior & marginal effects
- Techniques: feature importance, partial dependence plots

Explaining LLMs

- Techniques for transparency:
  - Feature attribution
  - Attention-based methods
  - Example-based explanations
  - Chain-of-Thought prompting

Implementation

Conceptual choices

- Focus on human likeness and clean design
- Use of LLMs for self-explanations
- Highlighting key words in prompts

Architectural Choices

- Python backend
- Simple web interface (HTML, CSS, JS)
- Tools: OLLAMA, Mistral model, NLTK

Demonstration

Conclusion

- Enhanced explainability of AI models
- Tools: AIX360, ELI5, InterpretML, Dalex2
- Techniques for LLM transparency
- Practical application in conversational agents

References

1. Trusted AI. “AI Explainability 360 (AIX360),” GitHub repository. Available: AIX360.
2. TeamHG-Memex. “ELI5,” GitHub repository. Available: ELI5.
3. ModelOriented. “Dalex2,” GitHub repository. Available: Dalex2.
4. Zhao et al. “Explainability for Large Language Models: A Survey,” arXiv. Available: arXiv.
5. Rheu et al. “Trust-Building Factors and Implications for Conversational Agent Design,” International Journal of Human-Computer Interaction, 2020. Available: DOI.