Explainable AI – Backgrounder

Explainable Artificial Intelligence (AI) increases the trust in AI solutions, enabling wider adoption and providing a foundation for AI governance.

INTRODUCTION

Humans have the ability to explain their decision-making process - logical thinking, observation, intuition, or experience. Basic machine learning algorithms such as decision trees can be understood by following the logic path that led to the decision. However, complex AI systems have traditionally been "black boxes" - even while knowing inputs and outputs, there was no understanding of the algorithms used to arrive at a decision. For humans to trust those decision, AI's rationale and algorithms need to be explainable.

EXPLAINABLE AI

Explainable AI (XAI) is an artificial intelligence that can describe its purpose, rationale and decision-making process in a way that humans can understand.

Explainable AI can be achieved it two ways:

- building the capability into the AI model from the beginning; or
- by using tools that describe how the decision was made after the fact.

Ideally, XAI would provide information on:

- inputs: how data was selected, how the training data differed from the production data, how it was examined for bias and how the bias was mitigated;
- process: which models were applied, why more weight was given to some over others, and how they were tested;
- outputs: how the inputs and models lead to the outputs, and how the outcomes were audited;
- performance: monitoring and reporting on how the system performed in the production environment.

WHY IS XAI IMPORTANT

The complexity and proprietary nature of some algorithms can obscure how they make decisions and could, therefore, mask harmful behavior. Bias in training data could have adverse effects on employment, immigration, or financial decisions; incorrect decisions can violate privacy, harm personal health or cause disastrous military mishaps.

Invalid AI decisions may lead to harm to humans, operational challenges for an organization, customer dissatisfaction, negative media attention, high employee turnover and regulatory scrutiny. When errors occur it is important to understand whether they were caused by invalid data, bugs in the software, incorrect models, or some combination of those elements. Even if the decision appears plausible, humans need to understand it for curiosity, peace of mind, legal compliance or for ethical reasons.

For all of these reasons, lack of transparency has a negative impact on trust and acceptance of AI. Rising legal and privacy aspects will also limit application of AI if decision-making cannot be explained. As artificial intelligence becomes increasingly prevalent, it is becoming more important to disclose how bias and the question of trust are being addressed.

Emerging Governance

The European Union's General Data Protection Regulation (GDPR) stipulates that when a human is subject to a decision made by an AI, the human has the right to receive an explanation, to challenge the decision, and to be guaranteed protection from profiling.

In January 2019, the European Commission's launched an Artificial Intelligence for European Union project with a goal to build the first European on-demand AI platform and to promote the development of explainable, verifiable AI.

The Canadian Government issued a Directive on Automated Decision-Making in April 2019 "to ensure that Automated Decision Systems are deployed in a manner that reduces risks to Canadians and federal institutions, and leads to more efficient, accurate, consistent, and interpretable decisions[1]." Within the Directive is an Algorithmic Impact Assessment (AIA) tool, designed to assess and mitigate the risks associated with deploying automated decision systems. The AIA classifies decisions made by AI according to their likely effect on the rights, well-being, or economic interests of individuals or communities, or ongoing sustainability of an ecosystem. Those deemed to have higher impact are subject to more rigorous explanation requirements.

[1] Section 4: Objectives and Expected Results, https://tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592

NEXT STEPS

The industry recognizes the importance of XAI and is taking steps toward its wider application:

- IBM's researchers proposed the concept of factsheets for AI services that would report on its fairness, robustness, explainability and lineage[2]. The researchers believe that standardizing and publicizing this information is key to building trust in AI services across the industry.
- The United States' Defense Advanced Research Projects Agency is researching "third-wave AI systems," where machines understand the context and environment in which they operate, and over time build underlying explanatory models[3].

Vendors are adding explanation capability to AI platforms in various ways:

- data science platforms that automatically generate model explanations in natural language;
- pattern visualizations for model analysis;
- tools for insights into neural network performance;
- transparent automated machine leaning models for financial decisions;
- tools that explain model findings and alert users to potential bias in data.

The efforts by solutions providers to address explainability, bias and compliance, combined with the sound regulations is a step towards creating a foundation of trust and setting a path to successful and widespread AI adoption.

Related Topics: AI, Augmented Analytics, Machine Learning

[2] Lineage means details of AI's development, deployment, and maintenance

[3] More information at https://www.darpa.mil/program/explainable-artificial-intelligence