Explainable Artificial Intelligence

Vanessa Keppeler
PwC, Germany

Matthias Lederer
Technical University of Applied Sciences Amberg-Weiden, Germany

Ulli Alexander Leucht
PwC, Germany

INTRODUCTION

Artificial intelligence (AI) can be used in almost all areas of a modern digital enterprise. While the very first AI systems were easy to interpret, increasingly opaque decision-making systems have emerged in recent years (Arrieta et al., 2019). This is largely due to the fact that their tremendous progress in performance has made them increasingly complex, making it difficult to understand how they come to a decision or outcome (Biran et al., 2017). AI systems are therefore often referred to as a 'black box' (Bauer et al., 2021). These complex and non-transparent models present a significant challenge to many companies when it comes to assuming responsibility and ensuring the traceability of decisions.

The explainability of AI thus represents one of the central barriers for the comprehensive use of the new technology. While there is widespread agreement on the basic requirement of explainability, the design of an appropriate explanation is rarely well defined and the definition of what 'explainable' means is less clear. There are different ways to formulate explanations, but it is not defined which formulation is considered appropriate to make AI explainable (Gilpin et al., 2019).

The comprehensibility and explainability of AI systems and their results is a basic prerequisite for the use and acceptance of the technology in many companies (Manikonda et al., 2020). In the research field of 'Explainable AI', the generation of explanations and the establishment of comprehensibility for AI systems is being researched intensively (Bauer et al. 2021). This includes all decisions being prepared or performed by highly complex AI models (Arya et al., 2019).

This contribution shows that there are numerous approaches for explanations, with varying relevance for different interest groups. Furthermore, the quality of explanations for artificial intelligence is described in more detail and evaluated with respect to completeness of an explanation as well as its interpretability.

BACKGROUND

Explanations

Explanations serve to make facts interpretable for human beings (Keil, 2006; Bechtel et al., 2005; Chater et al., 2006). Accordingly, an explanation is a mean or an instrument that helps people comprehend and understand decisions (Arrieta et al., 2019).

DOI: 10.4018/978-1-7998-9220-5.ch100

According to Karl Popper (Popper, 1935), an explanation serves to describe the cause of a decision by formulating its logical and causal relations. This is done by deductively deriving its laws and boundary conditions.

Popper attributes two essential components to an explanation. On the one hand that is a general hypothesis, e.g., 'everytime something happens, it has this same consequence', somewhat of a law for the case in question. The second component comprises particular statements, valid only for one specific occurrence, e.g., 'this happened' and 'this was the consequence in that case'. These case-specific statements represent the boundary conditions (Popper, 1935).

A causal explanation as such includes all components or factors related to a particular issue in a fully comprehensive way (Herman, 2019). However, for the explanation of a specific issue, people usually prefer short, selective explanations. They do not expect all causes to be fully included in an explanation, but rather that the most important ones are summarized (Molnar, 2020).

Furthermore, explanations are social interactions between the explainer and the receiver of the explanation (Confalonieri et al., 2020). Therefore, the social context has a great influence on the actual content of the explanation. Depending on the recipient, explanations are shaped differently (Van den Berg et al., 2020).

The quality of an explanation can be evaluated in two ways (Gilpin et al., 2019):

- **Interpretability**: The goal of interpretability is to provide an explanation for a decision that is as comprehensible as possible for humans (Doshi-Velez et al., 2017). The success of this goal depends on the perception, knowledge, and personal bias of the recipient (Miller, 2018). Therefore, for something to be interpretable, it must be described in a way that is simple enough for a person to understand, using vocabulary that makes sense to the counterpart (Gilpin et al., 2019).
- **Completeness**: The completeness of an explanation aims at describing an issue as precisely and accurately as possible. An explanation is more complete the better it makes an issue (e.g., a decision, a system, or an outcome) predictable in multiple situations (Gilpin et al., 2019).

There is a general trade-off between interpretability and completeness (Hoffmann et al., 2019). The most accurate explanations are not easy for people to interpret and, adversely, explanations that are particularly easy to interpret often do not offer extensive predictive power (Doshi-Velez et al., 2017).

Herman (2019) points out that the quality of an explanation for establishing interpretability therefore should not be assessed solely on the basis of human judgments and perceptions, because human judgments imply a strong and specific bias against simpler descriptions. In practice, this can lead to the presentation of simplified descriptions of complex systems to increase the confidence of recipients without them knowing the limits of the simplification (Doran, 2017).

To avoid this ethical dilemma, explanations should allow a trade-off between interpretability and completeness (Doshi-Velez et al., 2017). Depending on the recipient, an explanation should be able to be simplified. Vice versa, it should be able to be completed and expanded to include details, even though it then loses interpretability for humans. The evaluation of an explanation should be based on how it performs on the curve between maximum interpretability and maximum completeness (Doshi-Velez et al., 2017).

Related Content

Autonomous Last Mile Shuttle ISEAUTO for Education and Research
Raivo Sell, Mairo Leier, Anton Rassõlkinand Juhan-Peep Ernits (2020). *International Journal of Artificial Intelligence and Machine Learning (pp. 18-30).*
www.irma-international.org/article/autonomous-last-mile-shuttle-iseauto-for-education-and-research/249250

Adoption of Machine Learning With Adaptive Approach for Securing CPS
Rama Mercy Sam Sigamani (2022). *Research Anthology on Machine Learning Techniques, Methods, and Applications (pp. 1165-1192).*
www.irma-international.org/chapter/adoption-of-machine-learning-with-adaptive-approach-for-securing-cps/307505

Cluster Analysis as a Decision-Making Tool
Bindu Raniand Shri Kant (2023). *Encyclopedia of Data Science and Machine Learning (pp. 382-409).*
www.irma-international.org/chapter/cluster-analysis-as-a-decision-making-tool/317461

Opportunities and Challenges for Artificial Intelligence and Machine Learning Applications in the Finance Sector
Ruchi Sawwalakhe, Sooraj Aroraand T. P. Singh (2023). *Advanced Machine Learning Algorithms for Complex Financial Applications (pp. 1-17).*
www.irma-international.org/chapter/opportunities-and-challenges-for-artificial-intelligence-and-machine-learning-applications-in-the-finance-sector/317013

A Survey on Arabic Handwritten Script Recognition Systems
Soumia Djaghbellou, Abderraouf Bouziane, Abdelouahab Attiaand Zahid Akhtar (2021). *International Journal of Artificial Intelligence and Machine Learning (pp. 1-17).*
www.irma-international.org/article/a-survey-on-arabic-handwritten-script-recognition-systems/279276