AI4media

ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

ROADMAP ON AI TECHNOLOGIES & APPLICATIONS FOR THE MEDIA INDUSTRY

SECTION: "INTERPRETABLE AI"

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

info@ai4media.eu www.ai4media.eu

Authors	**Mara Graziani** (University of Applied Sciences of Western Switzerland - HES-SO Valais) **Henning Muller** (University of Applied Sciences of Western Switzerland - HES-SO Valais)

This report is part of the deliverable D2.3 - *"AI technologies and applications in media: State of Play, Foresight, and Research Directions"* of the AI4Media project.

You can site this report as follows:

F. Tsalakanidou et al., Deliverable 2.3 - AI technologies and applications in media: State of play, foresight, and research directions, AI4Media Project (Grant Agreement No 951911), 4 March 2022

This report was supported by European Union´s Horizon 2020 research and innovation programme under grant number 951911 - AI4Media (A European Excellence Centre for Media, Society and Democracy).

Copyright

Interpretable AI

Current status

The research field in ***AI interpretability*** has grown very quickly in the last decade. The literature on interpretability techniques counts (until 2020) more than 70,000 papers containing either "XAI", "explainability", or "interpretability"[1]. As Doshi-Velez and Kim argue[2], the rising interest in "***opening the black-box***" is motivated by the fact that evaluating the performance of complex machine learning models is an ill-posed problem, and that the sole model accuracy on the test is not sufficient to describe the model's inner functioning and the satisfaction of important desiderata. In applications where making a mistake would have strong implications on people's lives (e.g. credit allowance, insurance premiums, healthcare, etc.), the work on interpretable (or explainable) AI has emerged as a way to provide individuals with insights about automated decision-making. The main concept of explainable AI is illustrated in Figure 1.

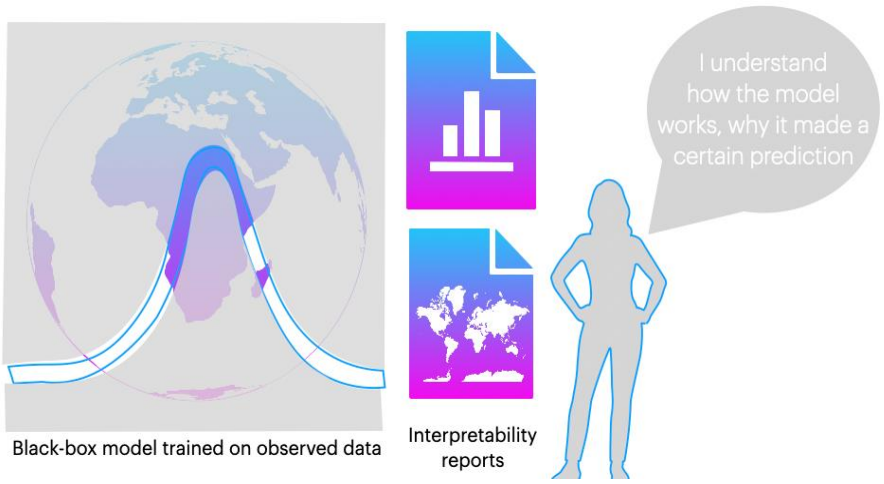

Figure 1: Explainable AI – the concept.

The EU's General Data Protection Regulation (GDPR)[3], in effect since May 2017, has officialised the need for ***reliability in addition to accuracy*** of the models, where reliability includes the generations of explanations that assign meaning to AI decision-making, improving the user's mental model of the automated process. Interpretability thus stands as a part of the social interaction between the AI system and its user. As we would expect bankers to explain why they rejected a loan, doctors to explain why they decided to discontinue treatment, and politicians to explain why they wanted to implement a certain policy, we would expect AI systems to justify their decision making if it impacts our lives. The GDPR forms, in this context, a "***right to be informed***", by claiming: (i) the right not to be subject to automated decision-making and safeguards enacted thereof (Article 22 and Recital 71); (ii) notification duties of data controllers (Articles 13–14 and Recitals 60–62); and (iii) the right to access (Article 15 and Recital 63). The

[1] From app.dimensions.ai, as accessed in August 2021.
[2] F. Doshi-Velez and B. Kim. "Towards a rigorous science of interpretable machine learning." arXiv preprint arXiv:1702.08608 (2017).
[3] General Data Protection Regulation (GDPR): https://eur-lex.europa.eu/eli/reg/2016/679/oj

information provided about the AI system must be meaningful to the individual confronted with the automated decision.

Some of the technical terms used to distinguish most of the current approaches to obtaining interpretable AI were introduced by Lipton[4]. Two important distinctions commonly adopted in the domain are that of (i) local vs. global explanations and (ii) built-in vs. post-hoc methods. ***Local explanations*** refer to explanations that are only true for a single input. ***Global explanations,*** on the contrary, explain the model behaviour for an entire set of inputs, e.g. all images of a single class in the dataset. ***Built-in methods*** introduce interpretability as one of the objectives of the model optimisation function. These methods are included in the more general notion of intelligible AI. An example is that of inherently interpretable models, e.g. linear regression, where the linear increase of a feature value corresponds to a proportional increase in the model output. ***Post-hoc methods*** are methods that generate explanations without requiring the retraining of the model parameters with interpretability constraints.

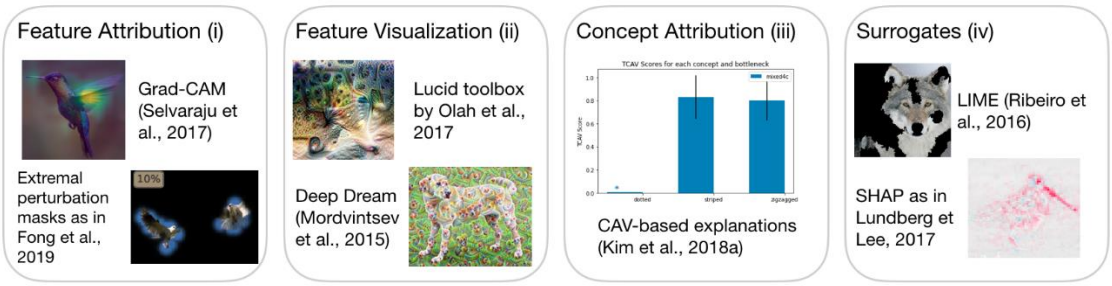

Figure 2: Post-hoc Interpretability approaches for AI models.[5]

Post-hoc approaches can be further grouped (as shown in Figure 2) depending on the form of the generated explanations into: (i) ***feature attribution methods***, that aim at identifying the input features that are the most relevant to the prediction; (ii) ***feature visualisation***, that aims at uncovering the patterns that are learned by intermediate layers and units; (iii) ***concept attribution***, that explains the model outcome in terms of high-level concepts and (iv) ***surrogate explanations***, namely those techniques that use a proxy model (generally simpler to interpret) to generate explanations. Two additional strategies are case-based and textual explanations. Another important distinction is between model-agnostic and model-dependent models. ***Model-agnostic methods*** do not need any access to the internal model's logic and/or state (e.g. model parameters), and only rely on the input and output pairs. They consider the model to be interpreted as a black box where only the output for a given input is observable. As a result, model-agnostic methods can be applied to all models. Perturbation methods such as occlusion and Local Interpretable Model-agnostic Explanations (LIME) are model-agnostic.

Most of these techniques still require large testing with users. Most often, human beings often do not need complete causal chains of explanation and may prefer a trustworthy account of

[4] Z. Lipton, "The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery." Queue 16.3 (2018): 31-57.

[5] Image taken from M. Graziani. Interpretability of Deep Learning for Medical Image Classification: Improved Understandability and Generalization. Diss. University of Geneva, 2021

understandable reasons expressed in clear and simple language. Collaborations should thus be built to develop types of human-computer interactions in ML that are more understandable to non-ML experts.

Research challenges

The main research challenges related to AI explainability can be summarised as follows:

- There is ***no ground truth of what constitutes a good explanation*** on real-world problems. While for controlled datasets it may be possible to identify the set of variables that are relevant to the data generation process, when dealing with the complexity of the data describing real-world phenomena such as the spreading of tweets and news, it is impossible to know a-priori the set of sufficient features to describe the phenomenon. Explanations, besides, may be differently understood by people with different backgrounds. Because of the subjectivity of the receivers of explanations, it is hard to define what an optimal explanation should be.
- The ***ability of understanding is highly subjective***. Computer scientists may understand information differently from experts in other domains. The analysis of the impact of subjectivity on the explanations is a major challenge. Anthropology should help understanding what characters, e.g. journalists, fact-checkers, content-creators, would define an "understandable" explanation.
- ***Interactive explainability*** techniques should be provided to allow humans to work together with AI systems. The explanations should then be related to the practices and the context. If a journalist were to use a fact-checking tool using AI, for example, to determine the reliability of a piece of information, then the fact-checker reliability should also be evaluated through explanations that are related to the practice and context of the journalism.
- Explainability should relate to ***eventual causation links*** between the features learned by the model and the features that are actually used to make the prediction. A convolutional neural network may be learning, for example, to extract visual features of texture, colour and shape from images depending on the architecture being used and how the filters, skip connections, and invariances are encoded in the model. Despite learning such features, only some of these may then be used to make the prediction. The aim of causal analysis is thus to evaluate not only *what features are being learned*, but also *how these features are used by the model* and *how changing such features would impact the classification decision*.

Societal and media industry drivers

Vignette: An interactive and explainable fact-checking tool for journalists

Glenn has been commissioned to verify the reliability of the viral spreading of information about the "*beginning of human cloning in some European countries as a means to promote the availability of organs for transplants*". He is strongly convinced that this information is false, being against the ethical rights and policies currently established by the Union. However, he can find confirmation of this news on multiple reliable sources. Puzzled, he decides to use a multimodal AI system with a content-retrieval module to verify the credibility of this information

by cross-matching all the existing evidence. The AI system retrieves and backtracks all the available information online about the news. Several content from yet un-regulated social media is filtered out and automatically reported as a non-cross-validated information on true-or-fake.socialmedia.com. As expected, the AI system clarifies that the information is a fake.

Relieved, Glenn starts digging about the AI outcome to uncover the reasons for this prediction. The system provides an interacting interface where it is possible to explore the information that was used to make the prediction in multiple ways. Some input images retrieved from online are highlighted in some regions pointing to the time of the information publication. Some checkered artifacts are pointed in several videos. Surprisingly, the model highlights as very important apparently unrelated documents in multiple languages that report power shortages at multiple media sites, including BBC, CNN and several European national televisions. At this point, Glenn interacts with the model to further illustrate the connection between the power shortages and the prediction of the news as fake. The AI system presents Glenn a cascading analysis of the events that influenced the prediction. The documents on the power shortage all document the same time as the actual electricity interruption. No major security issue was reported, so this information had been overlooked. The videos and articles about the beginning of human cloning, however, all reported a publication time within the power shortage. The clear signs of corruption in the images such as the checkered artifacts show that the videos were actually generated by other AI methods. Because of all of these reasons, the model concluded that it is fake news. Glenn now has more relevant questions to verify, namely whether there was a synchronised attack to major media industries to finally diffuse fake information again.

This was the first case of viral spreading of fake information again in the past 15 years, since the creation of true-or-fake.socialmedia.com and the use of explainable AI to detect misinformation.

Future trends for the media sector

Some of the opportunities concern the use of interpretable AI as a means to empower the possibilities of journalists, fact-checkers, film-makers, content providers, content creators, advertisers and other figures in the media. Interpretability may be used to evaluate on the basis of what reasons some content was predicted as fake. Disparities in the automated decision making due to bias towards gender, race and income will be highlighted by the explainability tools to promote equity and inclusivity. Interpretability may be used to identify bias and discrimination in models trained on incomplete and unbalanced datasets. For example, it may be used to guarantee that the tools for automatic casting in the film industry do not penalise non-white people. Similarly, it may be used to evaluate the presence of negative bias towards the recommendation of LGBTQ contents.

Goals for next 10 or 20 years

The future will see AI applications as part of a partnership with humans for empowering their capabilities. Knowledge must be collected on how humans acquire new information and represent their beliefs. Interpretable AI shall consider this knowledge to provide decision support. The systems shall be interactive, so that they may adapt to the user's needs and clarify any unclear explanations with further elaboration. Systems may directly point out societal risks

such as biases, discrimination and misinformation as soon as possible. The AI user of the future is informed at all times and builds trust in the machine by interacting with the system.

AI4media

ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

info@ai4media.eu www.ai4media.eu