Trustworthy Artificial Intelligence

Maximilian Becker

Vision and Fusion Laboratory
Institute for Anthropomatics
Karlsruhe Institute of Technology (KIT), Germany
maximilian.becker@kit.edu

Abstract

Trust or trustworthiness are hard to define. There are many aspects that can increase or decrease the trust in an Artificial Intelligence systems. This is why entities such as the High-level expert group on AI (HLEG) and the European commission's artificial intelligence act are putting forward guidelines and regulations demand trustworthiness and help to better define it. One aspect that can increase the trust in a system is to make the system more transparent. For AI systems this can be achieved through Explainable AI or XAI which has the goal to explain learning systems. This article will list some requirements from the HLEG and the European artificial intelligence act and will go further into transparency and how it can be achieved through explanations. At the end we will cover personalized explanations, how they could be achieved and how they could benefit users.

1 Introduction

Trust and trustworthiness are complex and not easy to define concepts. An organization that focuses on trustworthy artificial intelligence is the High-level expert group (or HLEG) on artificial intelligence[1]. The HLEG was appointed by the European Union to advise on their artificial intelligence strategy. They released an ethics guideline for trustworthy AI in which they define seven requirements for trustworthiness in AI systems[2]: 1) human agency and oversight, 2) technical robustness and safety, 3) privacy and data governance, 4) transparency, 5) diversity, non-discrimination and fairness, 6) environmental and societal well-being and 7) accountability. The HLEG has the aspiration to shape the EU's future approach to AI. With their ethics guideline they took their first step to make AI more trustworthy by listing their requirements. In this article we are going to focus on transparency.

The European Union also put forward a draft for a regulation called the artificial intelligence act which should enable the development and deployment of trustworthy AI. The new regulation is called the artificial intelligence act and will be covered in section 2. One way to increase the transparency of AI systems is through explainable AI or XAI. The goal of this field is to generate explanations for learning systems. Section 3 will give an overview over the field and how these explanations can look. Afterwards section 4 describes five different concrete approaches to explainability and give examples for each approach. Finally section 5 will look into personalizing these explanations, which means that the explanations are adapted to a users wants and needs, to make them more relevant for individual users.

2 Artificial Intelligence Act

The Artificial Intelligence Act or AI-Act for short is a draft for a regulation from the European commission from 2021. The full name is: Proposal for a Regulation laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts[4]. It is a legal framework similar to the GDPR [5] but written specifically for AI systems and

[1] High-level expert group on artificial intelligence, `https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai`

[2] Ethics guidelines for trustworthy AI, `https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai`

should lay the groundwork for trustworthy AI. The technologies falling under the new regulation are defined as [4]:

- Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning"
- "Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems"
- "Statistical approaches, Bayesian estimation, search and optimization methods.

The regulation puts forward transparency rules for AI systems used in these applications[4]:

- Interaction with humans
- Emotion detection
- Biometric identification
- Generation and manipulation of content such as Deep fakes

The regulation defines four risk levels: unacceptable risk, high risk, and low or minimal risk. Applications that fall under the first level are mostly prohibited. They are for example [4] systems that exploit vulnerabilities such as disabilities and are likely to cause physical or mental harm, real-time remote biometric identification in public places for law enforcement, or systems that use subliminal techniques beyond a persons consciousness and may cause physical or psychological harm. The focus of the regulation is on high risk applications for which it poses strict requirements. Under this category are [4] systems used as safety components in other systems as well as ones used in some critical areas such as biometric identification, management of critical infrastructure, education and vocational training [4]. The last two levels, low and minimal risk are not further defined and the implementation of the regulation is on a voluntary basis.

The regulation lists requirements for high risk systems like a risk management system[4], testing procedures, technical documentation, transparency rules and others. The transparency rules state that the systems operation is sufficiently transparent to enable users to interpret the systems output and use it appropriately[4]. These transparency criteria can be achieved through XAI. Explainability research focuses on the one hand on making the output of systems more interpretable and how to present these explanations to the user. On the other hand it focuses on explaining the inner workings of models in order to better understand the models, their scope and their boundaries which enables users to use the systems appropriately. So XAI technologies are perfectly suited to fulfill these transparency requirements.

3 Overview over XAI

Many modern learning approaches like deep neural networks are very powerful but they are black boxes to developers and users. This means that the models are very good at making predictions but it is not clear how they make their decisions. Explainable Artificial Intelligence or XAI has the goal to explain the decisions of learning systems. To do this there are generally 2 approaches: explain existing methods like deep neural networks or use inherently explainable approaches. Both approaches have their advantages and disadvantages. In the first case any model can be used and the explanation can be generated post-hoc. This means that no compromise on the used model necessary. However, these explanations often explain only part of the model or an approximation of it. So it can happen that the explanation is only some kind of artifact that is not really present in the model or data. The second approach uses models that are inherently explainable. This means that the model itself can be understood and not only an approximation or part of it. But there is often a trade-off between predictive ability of a model and its explainability. This means that an explainable model is in general less powerful than a black box model. To illustrate this, we can compare a deep neural network to a small decision tree. The neural network will make much better predictions but it is not clear how or why it makes these predictions. The decision tree on the other hand is completely intelligible because one can just check every node on the path that lead to a certain prediction but this comes at the cost of the models predictive power. A small decision tree will not be able to represent a complex classification problem.

Different explanation methods can be distinguished further[3]. An explanation can be global or local. Global explanations explain a whole model while local explanations only explain single predictions of the model. They can also be model agnostic or model specific. Model agnostic explanation methods can be used to explain any model while model specific explanations can only explain one or several specific models.

Another distinction is the dependence on training data. Explanations can be data dependent, which means that they need to be trained with data that is usually the training data of the model they should explain. They can also be data independent, which means that they only need access to the model itself or its predictions.

4 Approaches to XAI

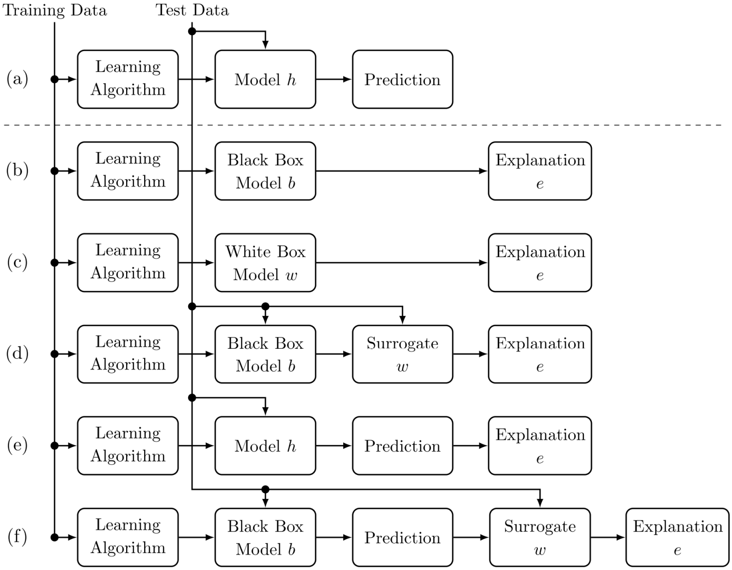

Figure 4.1: Default supervised learning and five different approaches to explainability[3] a) supervised machine learning, b) post-hoc explainability, c) white box model, d) global surrogate model, e) direct local explanation, f) local surrogate model

According to Burkart et al.[3] there are five different approaches to explainability for supervised machine learning (figure 4.1 (a)). The first approach is post-hoc explainability (figure 4.1 (b)). Here a black box model is trained on the training data and an explanation method is applied to the model afterwards. This is a global approach because the model as a whole gets explained. Post-hoc explanations have the advantage that any model can be used to make predictions which ensures a high prediction accuracy. An example for post-hoc explanations are partial dependence plots [7]. They visualize the dependence of the prediction on different features.

The second approach are white box models (figure 4.1 (c)). This approach uses a white box model which is a model that is inherently explainable. Because the whole model is understandable this is also a global approach. However these models can suffer from the afore mentioned trade-off between predictive power and explainability. Examples are decision trees and the explainable boosting machine[10].

The next approach are global surrogates (figure 4.1 (d)). A surrogate is a replacement for a black box model that is more explainable. At first a black box model is trained and with the black box model and the training data a second surrogate model is trained. The black box model is used for predictions and the surrogate model is used to generate explanations. Because the whole surrogate is explainable this is also a global approach. An advantage is that the prediction accuracy is conserved because a black box model is used to make the predictions. But a problem of this approach is that the surrogate model is only an approximation of the black box model so there will be a difference in the predictions of the two models, if they were identical the surrogate could be used as a white box model. This difference means that explanations generated with the surrogate only approximate the decisions made by the black box model. As an example decision trees can be used as a surrogate to approximate another model and explain it.

The fourth approach are direct local explanations (figure 4.1 (e)). Here a model is trained and used to make predictions. These predictions are then explained. Because only individual predictions and not the whole model are explained this is a local approach. An example are counterfactual explanations[14] which explain the decision for one instance by providing a second instance that leads to a different, desired prediction. So the counterfactual is another instance from the feature space that lies past a decision boundary and is ideally close to the original instance. The two instances or just the difference between them can then be used as the explanation because they represent the change in the feature space that leads to a different prediction. For example if a credit application got denied a counterfactual explanation could be that the application would have been accepted if the credit amount was 1000 less.

The last approach are local surrogates (figure 4.1 (f)). They are similar to global surrogates but as the name suggests they are local explanations. A black box model is trained and used to make predictions. A local surrogate model is then trained on samples from the area around the prediction and used to generate explanations. This surrogate does not represent the whole black box model but just the decisions in the vicinity of the instance of interest. An example are local interpretable model-agnostic explanations or LIME[12]. To generate the explanation points around the instance to be explained are sampled and then labeled using the black box model. A linear model is then trained with these samples under consideration of their distance from the original instance. This local model then represents the black box model's decisions in the vicinity of the instance and can be used to explain which features contributed more or less to the decision for the instance.

5 Making Explanations Personalized

Some research exists on what kind of explanations are suitable for which target groups. Different target groups like developers, domain experts and end users have different demands and levels of understanding and this must be considered when choosing or developing explanations[2].

A great way to improve XAI methods is to make explanations more personal and adapt them not only to groups of people but to the individual users. This would probably benefit end users the most as they are more interested in the decisions and their consequences to their lives than in understanding the model that made them. Personalized explanations would make the results more relevant to the user because they are adapted to their individual needs, requirements or preferences. This makes them easier to apply and may lead to more trust in the system.

We are going to look at ways to personalize counterfactual explanations (see section 4) from here on. A first step is to make them actionable. This means that only features that are easily changeable by the user are considered in the explanation, so for example the gender a person will not be regarded in the counterfactual instance. This is already done[1, 11] but does not really consider a persons preferences only world knowledge. A next step would be to not only exclude features but weigh the remaining ones. This would enable a system to incorporate user preferences in much more detail. So for example a user may be able to change her job or salary rather easily but changing her residence is really hard. Such preferences could be incorporated into counterfactual explanations using a weighted distance metric in the search process[9]. Another possibility to make counterfactual explanations more realistic is to consider interdependence between features. For example getting a better education takes time so this will result in the person getting older. It is often said that counterfactual explanations should be sparse[6, 8, 9, 13]. This means that as few features as possible are changed. Different people can however also have different preferences on whether it is better to change a single feature by a lot or multiple features a little. This can also be considered when personalizing counterfactuals.

All these options give the user very detailed options to personalize her explanations. The drawback of this is that each user has to take action and setup their personal preferences. This may deter users from using such a system. A solution could be to make the process interactive. At first a generic explanation is generated and afterwards the user can tell the system that a certain feature should be changed less or not at all. This process could be repeated until a satisfying explanation is found. The data gained in this process could also be used to learn a users preferences and use them to improve future explanations.

6 Summary

In this paper we looked at different aspects of making AI systems more trustworthy. At first the HLEG on AI and the European Artificial Intelligence Act were presented. Afterwards we focused on explainability of AI systems in general and on different approaches to it. At the end a research proposal for personalized explanations was presented.

References

[1] André Artelt and Barbara Hammer. "Convex optimization for actionable\ plausible counterfactual explanations". In: *arXiv e-prints* (2021), arXiv–2105.

[2] Alejandro Barredo Arrieta et al. "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI". In: *Information Fusion* 58 (Dec. 2019). DOI: `10.1016/j.inffus.2019.12.012`.

[3] Nadia Burkart and Marco F Huber. "A survey on the explainability of supervised machine learning". In: *Journal of Artificial Intelligence Research* 70 (2021), pp. 245–317.

[4] European Commission. *Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*. URL: `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206`.

[5] European Commission. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. URL: `https://eur-lex.europa.eu/eli/reg/2016/679/oj`.

[6] Susanne Dandl et al. "Multi-Objective Counterfactual Explanations". In: *Lecture Notes in Computer Science* (2020), pp. 448–469. ISSN: 1611-3349. DOI: `10.1007/978-3-030-58112-1_31`. URL: `http://dx.doi.org/10.1007/978-3-030-58112-1_31`.

[7] Jerome H Friedman. "Greedy function approximation: a gradient boosting machine". In: *Annals of statistics* (2001), pp. 1189–1232.

[8] Thibault Laugel et al. "Inverse Classification for Comparison-based Interpretability in Machine Learning". In: *stat* 1050 (2017), p. 22.

[9] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. “Explaining machine learning classifiers through diverse counterfactual explanations”. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 607–617.

[10] Harsha Nori et al. “InterpretML: A Unified Framework for Machine Learning Interpretability”. In: *arXiv preprint arXiv:1909.09223* (2019).

[11] Rafael Poyiadzi et al. “FACE”. In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (Feb. 2020). DOI: `10.1145/3375627.3375850`. URL: `http://dx.doi.org/10.1145/3375627.3375850`.

[12] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “"Why Should I Trust You?": Explaining the Predictions of Any Classifier”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. 2016, pp. 1135–1144.

[13] Arnaud Van Looveren and Janis Klaise. “Interpretable Counterfactual Explanations Guided by Prototypes”. In: *Age* 46 (), p. 46.

[14] Sandra Wachter, Brent Mittelstadt, and Chris Russell. “Counterfactual explanations without opening the black box: Automated decisions and the GDPR”. In: *Harv. JL & Tech.* 31 (2017), p. 841.