Helping Your Future Self Understand the Explanations You Generated Today

Once you get an explanation technique up and running, it's tempting to cash in on your hard work and generate as many explanations as you can right away. Unfortunately, you often find yourself squinting at many charts in an old Jupyter notebook or trying to explain the context for the explanation you sent along to colleagues a few months ago.

To save yourself future frustration, we've found it useful to always embed the following information in your explanations:

1. Exact technique and parameters used, e.g., was it SHAP or Captum's sampled Shapley?
2. Model version, training configuration, hyperparameter values, and dataset version, to be able to trace the source of the explanation.
3. Timestamp the explanation was generated, which is useful for knowing if the explanation is stale.
4. Input and inference values. You would be surprised how few techniques include this information in their visualizations.

Assuming Causality

Very few, if any, explanation techniques are able to establish causality in any sufficiently complex model. Techniques can only describe correlations between what influenced the model and the prediction. For example, Integrated Gradients may highlight a single pixel as highly influential in the model's prediction, but the technique does not guarantee that the pixel caused (even in part) the prediction.

At odds with explanation techniques' ability to provide correlations is the strong human desire to explain consequences due to causality. Causality is an important part of storytelling and narratives, and often you will find that consumers try to fit an explanation into a broader narrative to justify, attack, or just comprehend a model's actions. It is very difficult to work around this need for causality, and you will not get far trying to change your consumer's instinctive behavior. Instead, there are two strategies you can use to mitigate the tendency to fall back to causative descriptions:

- Language matters. Whenever introducing an explanation, whether with text, verbally, or in a presentation, be careful to not introduce or imply causation. This can be very difficult! For example, with feature attributions, it is tempting to say a particular feature caused the model to behave a certain way. Instead, try to use words like "influence" or "suggest."

- Avoid "this-then-that" narratives. Often explanations, with good intentions, try to present a logical flow of information and narrative. "This is the input to the model, then the model generated this prediction, here is the explanation" is a common narrative. Unfortunately, this narrative also implies a causal chain of reasoning from inputs to explanations. Instead, you may want to try inverting this narrative: "The model gave this prediction, which is explained by X. Additionally, here are the inputs."

Overfitting Intent to a Model

When given a sufficiently compelling explanation, consumers are tempted to extrapolate from the explanation to concepts learned by the model. Except for those focused on concepts, e.g., TCAVs, it is difficult to say that most explanation techniques are able to reveal semantic concepts the model has learned. In our earlier example of an image classifier given a prediction for a photo of a cat, it is accurate to say what pixels influenced the prediction, but it is not accurate to reach further and say, "Now we know the model has learned how to recognize cat ears." It certainly could have, but a pixel attribution technique gives explanations based on the pixels in the image, not the semantic concepts related to those pixels.

To avoid this overfitting, make explanations clear and constrained.

Overreaching for Additional Explanations

Once given a sufficient explanation, it is not unusual for ML consumers to reach for another explanation technique to augment their understanding. However, these techniques, even if good on their own, may not actually increase the power of the original explanation. For example, a common reach is for a user who has received a feature attribution explanation to try and find a counterfactual explanation to prove the validity of the feature attribution by finding a prediction for a data sample with a different value for the most influential feature and a different predicted class. The consumer may then declare this proves the influence of the top-ranked feature. While the counterfactual can enhance our understanding of the model's behavior, and even back-and-forth between looking at feature attributions and corresponding counterfactuals can tell us about different facets of the model, it is important to not treat each explanation as a validation of the other. Each explanation has its own gotchas and nuances that constrain what they can tell us about the model. Together, they may widen our understanding of the model, but may not necessarily deepen our understanding of one particular aspect.

Preventing explanation overreach is difficult because users often take matters into their own hands to find new explanations. Strongly discouraging this behavior rarely works in practice either, as the ML consumers will genuinely believe they are proactively contributing to the overall quality of the Explainable AI. Instead, to prevent

5.3.9.1 Human-Artificial Intelligence System

A human-AI system includes both computational elements and a human user who must come together to accomplish a purpose.

5.4 What is Explainable AI?

Explainability is a concept that stands at the crossroads of numerous fields of active AI research, with an emphasis on the following domains.

5.4.1 Fairness

Can we verify that choices taken by an AI system were done consistently?

5.4.2 Causality

Can one learn a system from facts that not only makes the right predictions but also offers some understanding of the core events?

5.4.3 Safety

Can we have confidence in the reliability of our AI system without recognizing how it makes its presumptions?

5.4.4 Bias

How can we be confident that the AI system hasn't picked up a distorted view of the world due to flaws in the training data or objective function?

5.4.5 Transparency

Everyone has a right to be informed about changes that influence us in ways, formats, and languages that we comprehend.

An XAI, also known as a “Transparent AI” or “Interpretable AI”, is an AI whose activities are simple for humans to comprehend and evaluate. A civic privilege to explain can be implemented using XAI.

5.5 Need for Transparency and Trust in AI

The black box AI systems have found their way into many of today’s modern implementations. Transparency and explainability are not critical requirements for machine learning models used as long as the overall efficiency of these systems is adequate. Even if these systems fail, the implications are unexceptional. As a result, the necessities for trust and openness in these types of AI systems are relatively low. The scenario is different in safety–critical applications. In this case, the opaqueness of ML techniques may be a restricting or indeed rejecting component. Particularly when a single misjudgment can endanger human life and health or lead to significant revenue damages, depending on an information system with unintelligible logic will not be an alternative. This lack of transparency is among the causes why the application of machine learning to areas such as healthcare is extremely careful than its application in the consumer, electronic commerce, or media industries.

5.6 The Black Box Deep Learning Models

The method of developing interpretations for AI system behavior will vary based on the type of ML techniques used: techniques that produce implicitly decipherable models vs deep learning algorithms that are intricate information and understanding methods and produce models that are implicitly indecipherable to actual users.

ML techniques such as Bayesian classifiers, decision trees, spare linear models, and additive models produce decipherable models in the sense that model components can indeed be instantly examined to comprehend the model’s inferences. These technique makes use of relatively small internals, and also provide visibility and traceability in their decision-making. As long as the model is precise for the classification process, these strategies offer awareness of the AI system’s decision-making.

Deep learning algorithms, one on either side, are a class of machine learning technique that sacrifices clarity and interpretability for the predictability. These techniques are now used to create applications such as consumer behavioral forecasting associated with high inputs, voice recognition, natural language processing, and computer vision.

The lack of transparency and understandability in the Deep Learning Algorithms makes them a black box. The black box model is a model which performs its predictions on its own without explaining anything for humans to understand.

The Black Box Problem occurs when artificially intelligent processor architectures are vague.

graceful feedback to the user when a user-error has occurred. On the other hand, as we know from usability research (Norman 2013) that the human-machine interface is also crucial for preventing the user from both conscious mistakes and unconscious slips during their interaction with the system.

11.3 Explicability—An Ethical Principle for Trustworthy AI

Compliance with the law is only one of the three components of trustworthy AI. Another is adherence to ethical principles and values, of which four are explicitly named by the High-Level Expert Group on AI: three traditional bioethics principles (human autonomy, prevention of harm, and fairness), which are in turn based on those described in the Charter of Fundamental Rights of the European Union (European Parliament, the Council and the Commission 2012), and a fourth: *explicability* (European Commission, Directorate-General for Communications Networks, Content and Technology 2019).

Explicability is a new ethics principle specifically relating to AI. It relates to the tendency for AI systems to act on the basis of complex internal processes that are invisible and/or unintelligible to humans (Floridi et al. 2018), rendering their decision-making processes difficult to understand, interpret, and explain (Holzinger et al. 2017). These are crucial issues for trustworthiness, validation, and acceptance of AI (Ziefle et al. 2013). According to Floridi et al. (2018), explicability recognizes the need to understand and hold to account the decision-making processes of AI.

To address the challenge of explicability, the field of *explainable AI* (XAI) research strives to provide insights into how a given AI model works and why it generates a particular result (Holzinger et al. 2018; Longo et al. 2020). There is a jumble of terms related to this concept in the XAI literature: with the terms explainability and interpretability often being used interchangeably (Zhou et al. 2021). Moreover, a variety of terms, including *transparency*, *accountability*, *intelligibility*, *understandability*, and *interpretability*, *comprehensibility* are used, sometimes interchangeably, sometimes with subtle differences in meaning that vary according to author. Other times, these terms are used without defining their specific meaning, or with one same term used for different meanings, or many different terms all referring to the same concept (Lipton 2018).

Gilpin et al. (2018) describe the concept of explainability as a combination of *interpretability* and *fidelity*, both of which are needed to achieve explainability. Here, interpretability refers to how understandable an explanation is for a human, and fidelity describes how accurately an explanation depicts the behavior of the AI model over the entire feature space. However, this often entails a trade-off between these two qualities, whereby it is difficult to simultaneously achieve both high interpretability and high fidelity: The most comprehensive explanation may not be easily interpreted by a human, and an intuitive explanation may not be sufficiently complete in its coverage of other usage scenarios (Gilpin et al. 2018). To reach optimal explainability, it is, therefore, necessary to assess the relative importance of each of these explainability properties in a specific application context.

Miller (Miller 2019) states that we know from social sciences that usually "*people ask for 'everyday' explanations of why specific events occur, rather than explanations for general scientific phenomena*" and he argues that this holds also in the context of Artificial Intelligence (Miller 2019). To be useful, any explanation must fit the tasks and goals of the receiver of this explanation. Therefore, for an efficient and effective explanation component in an AI system, it is crucial to take into account **who** uses **which** type of AI-solution for **what** purpose, and **how** the human-AI interface is designed (Müller et al. 2022).

11.4 User-Centered Approach to Trustworthy AI

For achieving explainability, as a precondition to trustworthiness, it is critical to develop a profound and comprehensive understanding of the purpose and context of the AI application in question. This includes detailed knowledge of the stakeholders who need to understand and interpret the results provided. With respect to this deep understanding of stakeholders, the article 9 of the Artificial Intelligence Act mandates that *"due consideration shall be given to the technical knowledge, experience, education, training to be expected by the user and the environment in which the system is intended to be used"* (European Commission 2021). To this end, the following section describes methodologies for generating the rich stakeholder profiles necessary for meeting these requirements.

11.4.1 Stakeholder Analysis and Personas for AI

To achieve the aforementioned requirements for trustworthy AI, it is necessary to focus on users and use-cases throughout the conception, scoping, and implementation stages of AI application development. For traditional computer applications, such a user-centered approach (Holzinger et al. 2005) has gradually been adopted over the past four decades. There is a large set of proven tools and methodologies available for the user-/human-centered design of conventional computer systems (Vredenburg et al. 2002). However, due to the specific characteristics of AI systems, many of these existing HCI tools and methods will need to be adapted and extended to effectively support their human-centered design and development (Xu et al. 2021).

One of the existing methods successfully applied in user-centered design of conventional computer applications is that of *Personas*. This method was introduced for user-centered interaction design by Alan Cooper in 1999 (Cooper and Saffo 1999). Personas are hypothetical user archetypes that help designers and developers to empathize with the target users, to focus on the needs and goals of these users deployment of XAI technologies. Future research should focus on developing ethical guidelines, standards, and frameworks for XAI that promote fairness, transparency, accountability, and privacy in AI-driven decision-making processes [4].

Advancements in explainable AI research hold great promise for enhancing transparency, interpretability, and trust in AI systems across various applications and domains. By providing transparent explanations for AI-driven decisions, XAI enables stakeholders to understand, validate, and trust AI models, leading to better decision-making processes, improved user experiences, and enhanced societal impact. As the field continues to evolve, interdisciplinary collaboration, regulatory alignment, and stakeholder engagement will drive the development and adoption of XAI, ultimately shaping the future of artificial intelligence and society [5, 6, 7].

5.2 Ethical and Regulatory Considerations

As artificial intelligence (AI) technologies continue to advance and permeate various aspects of society, ethical and regulatory considerations have become increasingly important. AI systems have the potential to bring about significant benefits, but they also raise complex ethical dilemmas and regulatory challenges. This chapter explores the ethical and regulatory considerations surrounding AI, examining key issues, guidelines, and frameworks aimed at promoting responsible AI deployment and mitigating potential risks [1].

5.2.1 Ethical considerations (Figure 5.1)

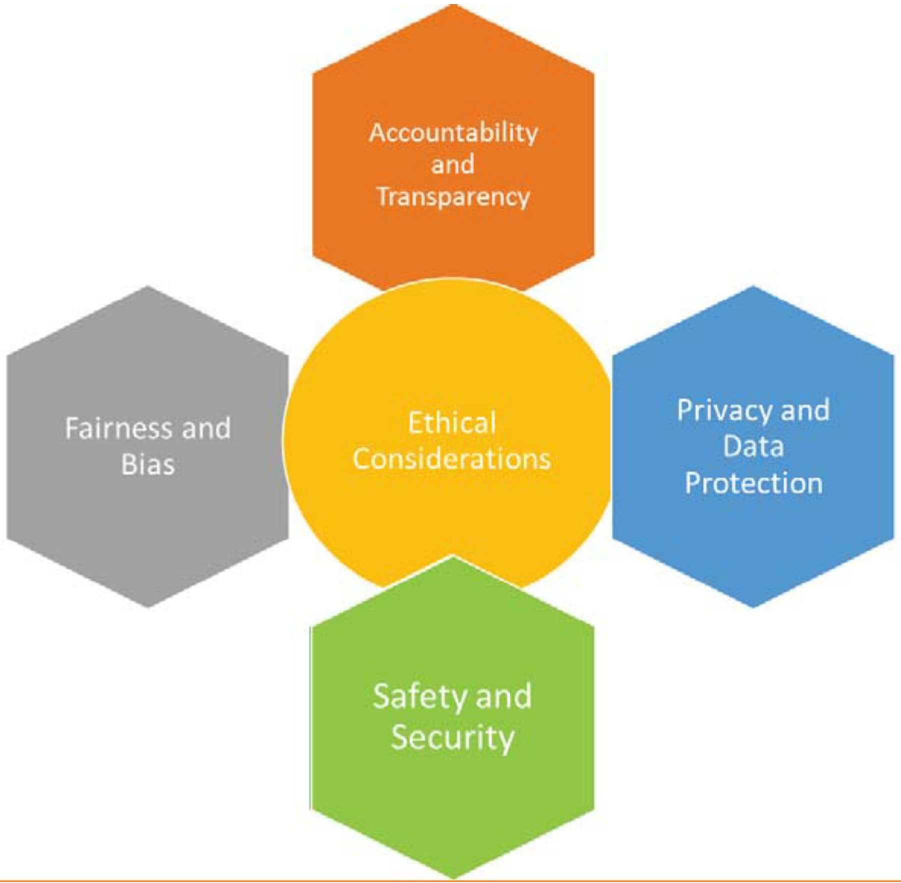

Figure 5.1: Ethical considerations.

Fairness and bias: AI systems can inadvertently perpetuate biases present in the data used for training, leading to unfair treatment and discrimination against certain groups. Addressing fairness and bias in AI requires careful attention to data collection, algorithm design, and evaluation methods to mitigate biases and ensure equitable outcomes for all individuals. Ethical considerations also extend to the allocation of resources, opportunities, and benefits generated by AI systems, ensuring that they are distributed fairly and transparently across diverse populations [7].

Accountability and transparency: AI systems operate as black boxes, making it challenging to understand how they arrive at their decisions. Ensuring accountability and transparency in AI requires mechanisms for explaining and justifying AI-driven decisions to stakeholders, enabling them to understand, validate, and trust AI systems. Ethical considerations also include establishing clear lines of responsibility and accountability for AI systems, delineating roles and obligations for developers, operators, and users to promote responsible AI deployment and usage [7].

Privacy and data protection: AI systems rely on vast amounts of data for training and decision making, raising concerns about privacy, consent, and data protection. Protecting privacy and data rights in AI requires robust data governance frameworks, encryption techniques, and access controls to safeguard sensitive information from unauthorized access or misuse. Ethical considerations also include respecting individuals' autonomy and privacy preferences, ensuring transparency and informed consent for data collection, storage, and usage in AI applications [7].

Safety and security: AI systems have the potential to pose risks to safety and security if deployed without adequate safeguards and risk mitigation strategies. Ensuring safety and security in AI requires rigorous testing, validation, and certification processes to assess AI systems' reliability, robustness, and resilience to adversarial attacks. Ethical considerations also include designing AI systems with fail-safe mechanisms, ethical AI principles, and human oversight to prevent unintended consequences and ensure responsible AI deployment in safety-critical domains such as healthcare, transportation, and defense [7].

5.2.2 Regulatory considerations (Figure 5.2)