Introduction to Machine Learning

Module 5: Interpretability and Ethics

Tobias Rebholz

University of Tübingen

Fall 2024, SMiP Workshop

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Explainable AI

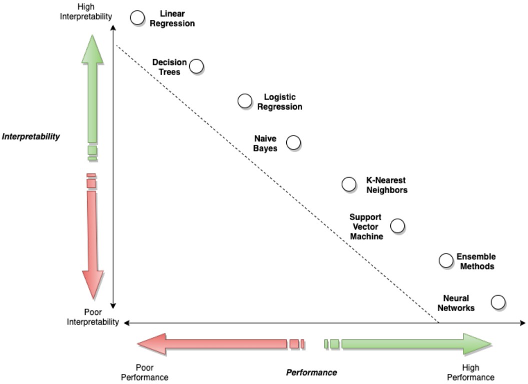
Explainable AI

In the history of science and technology, the engineering artifacts have almost always preceded the theoretical understanding

(Yann LeCun, Turing Award Winner)

Model Interpretability

- Problem: More powerful models are less understandable



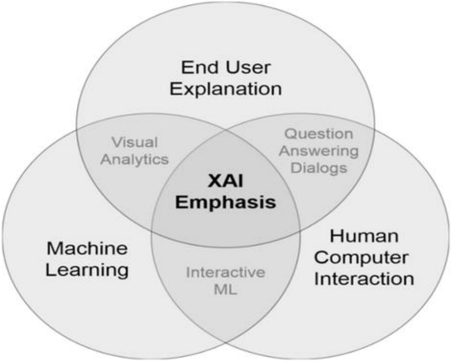
(https://docs.aws.amazon.com/whitepapers/latest/model-explainability-aws-ai-ml/interpretability-versus-explainability.html)

Model Interpretability

- **Black box problem:** ML algorithms, and particularly Deep Learning methods, solve problems in inscrutable ways
  - Partly because some of them can refine themselves autonomously and with an idiosyncrasy beyond the scope of human comprehension
- This is especially worrisome when the algorithm is making decisions with real-world consequences for human well-being, such as:
  - Determining whether a blip on a scan might potentially be cancerous
  - Applying the brakes in an autonomous vehicle
  - Granting a loan to buy a house
  - . . .

Explainable AI

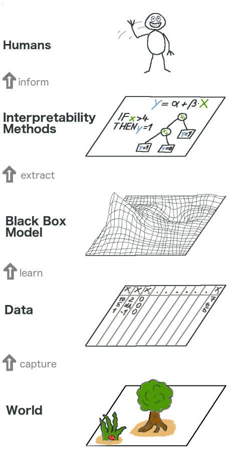
- **Explainable AI (XAI):** Subfield of ML that aims to address the black box problem by attempting to increase the interpretability, transparency, and ultimately also fairness of such methods



(Dağlarli, 2019, Figure 1)

Explainable AI

- The XAI process:



(Molnar, 2022, Figure 6.1)

Explainable AI vs. Interpretable ML

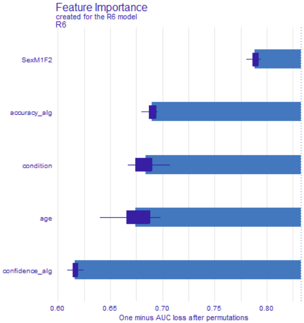
- **Explainable AI (XAI):** Methods and techniques used to make complex and opaque ML models more understandable to humans
  - This includes post-hoc explanations of models that are not inherently interpretable (e.g., a random forest consisting of many deep trees).
- **Interpretable ML (IML):** Developing models that are inherently understandable to humans
  - I.e., the inner workings of the model and how it makes decisions can be directly understood without additional tools or techniques (e.g., a single decision tree)
- Since we covered interpretable models at the beginning of the workshop, we will focus primarily on XAI techniques in this final module
  - However, we will use both terms (XAI/IML) **interchangeably**, referring to their common goal of making the predictions of ML methods more transparent

Model Agnosticity

- Some interpretation methods are **model-specific** (e.g., saliency maps for CNNs; see below)
  - I.e., they can only be applied to a certain model (family)
- Problem: ML requires **benachmarking**
  - Typically, not just one, but many types of ML models are compared against each other in their performance to solve a specific prediction task
  - For this reason, and for the sake of brevity, we will mainly discuss model-agnostic XAI/IML techniques in this module
- **Model-agnostic** interpretation methods: Separating the explanations from the ML model (i.e., can be applied to any ML model)
  - Freedom to use any model, essentially including the most powerful ones
- We already know an important model-agnostic XAI technique from Module 2:
  - See next slide

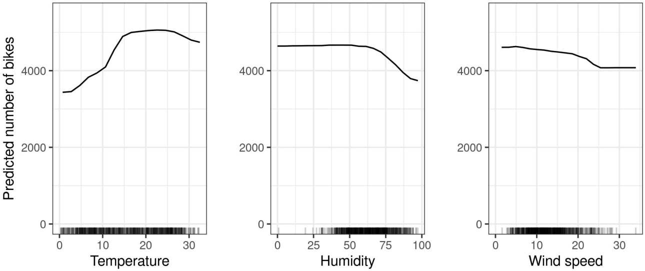
Important XAI/IML Techniques

- **Permutation feature importance:** Measuring the increase in the prediction error of the (random forest) model after randomly permuting the feature's values
  - Idea: Breaking the original relationship between a specific feature and the target



Important XAI/IML Techniques

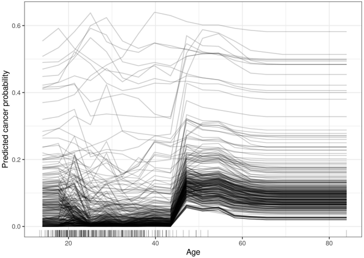
- **Partial Dependence Plots (PDPs):** Visualizing the relationship between a feature and the predicted outcome, averaging out the effects of all other features
  - I.e., showing the so-called **“marginal effect”** of a feature on a target
  - The relationship between a feature and the target can be linear, monotonic, or anything else (i.e., much more complex)



(Molnar, 2022, Figure 8.1)

Important XAI/IML Techniques

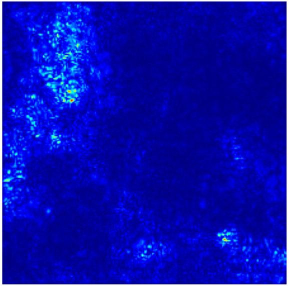
- **Individual Conditional Expectation (ICE) plots:** Visualizing how changes in a single feature affect the predictions of a ML model for (all) individual instances
  - Allows for an assessment of the **heterogeneity** of marginal effects across instances in a data set
  - In this sense, ICE plots are an extension of PDPs



(Molnar, 2022, Figure 9.1)

Important XAI/IML Techniques

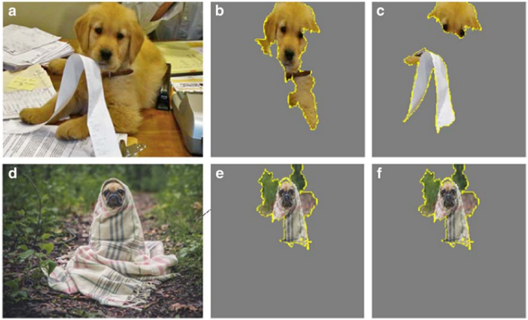
- **Saliency maps:** Highlighting regions in the input data (e.g., image pixels) that are most influential in the ML model's prediction
  - E.g., coloring pixels according to their contribution to the classification:



(Molnar, 2022, Figure 10.8)

Important XAI/IML Techniques

- **Local Interpretable Model-agnostic Explanations (LIME):** Highlighting the superpixel areas that are most important for an image classification
  - LIME-explanations for Inceptionv3's classifications:
    - b: golden retriever; c: toilet tissue
    - e: bath towel; f: three-toed sloth



Excurse: Further Important XAI/IML Techniques

- General idea behind LIME (Ribeiro et al., 2016): Explaining individual predictions by approximating the complex model (e.g., neural network) locally with a simpler, more interpretable so-called **"surrogate model"** (e.g., linear regression)
- **SHapley Additive exPlanations** (SHAP; Lundberg and Lee, 2017): Providing a measure for the contribution of each feature to the prediction
  - Cf. permutation feature importance
  - But: Based on the game theoretically optimal "Shapley values"
    - How to fairly distribute the "payout" (= prediction) among the "players" (= features)?
- **Counterfactual explanations:** Describing the minimal changes to input features that would alter the model's output (i.e., achieving a desired prediction)
  - Goal: Providing actionable insights about model sensitivity
- ...

XAI/IML in R

```
"Not discussed!"
```

- If you want to learn more about these methods, please refer to the list of references on the last slide!

Artificial Cognition

Artificial Cognition

- Computer scientists explain artificial intelligence primarily by tinkering under the hood of the black box, including attempts such as:
  - Generating explanations through more interpretable architectures
  - Introducing a second AI that examines another AI's decision and learns to generate explanations for it
- Problems:
  - Most XAI techniques are merely **correlational** in nature: Estimating which features the model cares about by displaying hypothetical predictions based on previously learned correlations
    - This is not bad per se: E.g., we can see when a model may be unfairly over-indexing on **protected attributes**, like gender or race (discussed in more detail later!)
  - **Malleable introspection:** Similar to asking humans how they made a decision, AI systems generate explanations that may not be the true (causal) explanation for their behavior

Artificial Cognition

- The human mind is also a black box!
  - Cognitive psychology: A science of behavior that works without opening its black box
- **Artificial Cognition** (Ritter et al., 2017; Taylor and Taylor, 2021): Translating the methods and rigor of cognitive psychology to the study of artificial black boxes to enable explainability
  - Satisfactory explanations: Causal, rather than correlative
    - Require experimental attempts of **falsification**
  - Subfield of the so-called "Machine Behavior" movement toward XAI
    - Interpretation methods that do not rely on AIs trying to explain themselves

Artificial Cognition

- Procedure: Using **behavioral experiments** to infer the properties of invisible artificial mental processes
  - E.g., by asking questions like: What if the input was a little bit different, would the output of the model be different as well?
  - Cf. approaching the black box problem for the human mind
- Incl. identifying **boundary conditions**: Being able to explain when a behavior occurs implies that we should also be able to account for when the behavior stops occurring
- Due to huge differences between ML model architectures, in contrast to human brain architectures, more akin to the psychology of **individual differences**
  - I.e., instead of central tendencies, deviations from the mean represent true data

Ethical Considerations

Model Bias

- The parameters of a fitted ML model reflect the truth derived from the training data, not necessarily the real world!
- Problems:
  - **Discrepancies** between the training data and the real world are inevitable
    - This can create several vulnerabilities for stakeholders (i.e., users and other parties affected by a model's outcomes)
    - E.g., training datasets that are biased against a particular group, such as gender or race, will yield model predictions that are biased against that particular group
  - Without a **causal** model, it is challenging to distinguish between direct discrimination (e.g., women earning less because they are women) and indirect discrimination (e.g., women earning less because they choose lower-income professions)

Model Fairness

- A trained ML model is deemed **fair** if it does not discriminate against protected subgroups
  - Fairness is context-dependent and involves normative consensus
  - Transparent documentation on model performance and intended use cases is crucial for assessing model fairness
- Fairness is particularly relevant for psychological assessments using ML, but potential remedies exist:
  - Avoiding to directly includ protected attributes (e.g., gender) in the model to prevent it from using these information
    - No panacea: Flexible ML models can **infer** group membership on the basis of available features that are related to the protected attribute
    - E.g., inferring socio-economic status from ZIP code (cf. imputation methods)
  - Explicitly evaluating fairness by comparing predictions for different values of the protected attribute using XAI techniques (e.g., feature importance or PDPs/ICE)

Model Fairness

- Challenge: Fairness and ethics are flexibly interpreted
  - Cultural differences: Influence perceptions of AI ethics, especially in high-stakes scenarios
    - E.g., car accidents of autonomous vehicles
  - Perspectives: Can shift based on stakeholders' roles
    - E.g., passenger vs. pedestrian in the context of autonomous driving
- XAI-based explanations will only be satisfactory if they enable stakeholders to judge whether the decision was appropriate in that particular situation

Model Accountability

- ML models should be accompanied by clear documentations, detailing their performance characteristics, limitations, and intended use cases
  - Transparency helps stakeholders understand the decision-making process and trust the model's outputs
- Selected accountability mechanisms:
  - Establishing **frameworks** to assess ethical operation of ML systems
  - Ensuring that shareholders (i.e., organizations that deploy ML models) can be held **responsible** for the outcomes
  - Continuous **monitoring** and evaluation of deployed models to detect and mitigate biases over time
  - Ensuring that data used for training ML models respect users' privacy and complies with relevant **regulations** (e.g., GDPR)
    - Incl. implementing data anonymization and secure data handling practices to protect sensitive information
  - Addressing **vulnerabilities** in ML models that could be exploited maliciously
    - Incl. robust defenses against adversarial attacks
  - Assessing the broader **societal implications** of deploying ML models
    - Incl. potential impacts on employment, social equity, and human rights
  - **Involving diverse groups** in the development/deployment to ensure that the perspectives and needs of different communities are considered

Model Trust

- Psychological factors that were found to influence people's trust in technology:
  - Providing clear, understandable information about—particularly black box—models' judgment and decision-making processes
  - Consistency and predictability of a model's behavior (e.g., in terms of performance)
  - User-centered design: E.g., (conversational) user interfaces that facilitate access and understanding
  - Users' prior experiences with similar technologies and their individual differences (e.g., tech-savviness, cognitive style)
  - . . .

Model Trust

- Measurement challenges:
  - Trust is **not static** but evolves over time based on user interactions with specific technology
    - More longitudinal studies and repeated measures designs are needed to understand how trust develops and changes
    - Continuous user feedback and iterative design processes can help build and maintain trust over time
  - Trust levels can vary greatly depending on the **context** of use and cultural background of the stakeholders
    - Cross-cultural studies are important to identify and account for these differences in trust
  - Self-reported measures of trust in technology should be complemented by behavioral data
    - Trust can also be inferred from user behavior, such as frequency of use or adherence to recommendations

Summary

Summary

- **XAI/IML:** Attempts to open the black box of ML-based AI technology
  - Vs. **Artificial Cognition**: Cognitive psychological approach (i.e., experimentation/falsification) to explain black box algorithmic behavior
- **Ethical considerations:** Gain importance in light of the increasing implementations of black box systems in the digital ecosystem
  - E.g., fairness, limitations, potential biases, transparency, accountability, privacy, security, trust, ...

Further Readings

1. **XAI:** Henninger, M., Debelak, R., Rothacher, Y., & Strobl, C. (2023). Interpretable machine learning for psychological research: Opportunities and pitfalls. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000560
2. **IML:** Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable* (2nd ed.). Christoph Molnar. https://christophm.github.io/interpretable-ml-book/
3. I.a., **ethics**: Van Dis, E. A. M., Bollen, J., Zuidema, W., Van Rooij, R., & Bockting, C. L. (2023). ChatGPT: Five priorities for research. *Nature*, *614*(7947), 224–226. https://doi.org/10.1038/d41586-023-00288-7