Explainability and Interpretability in AI Systems

Stanford CS221
Embedded Ethics Lecture, Week 9
Myra Deng, Veronica Rivera

Learning objectives

- Understand the differences between explainability and interpretability in AI systems
- Discuss trade-offs between simpler logic-based systems (more explainable/interpretable) and complex ML systems (less explainable/interpretable but often more performant)
- Describe emerging best practices for designing explainable and interpretable AI systems
- Highlight current research directions in explainability and interpretability

Explainability

Refers to the ability to retain human intellectual oversight over AI systems. Typically focused on making **decisions** made by an AI system understandable and transparent

"Can the model provide human-understandable explanations or justifications for its predictions or decisions?"

https://en.wikipedia.org/wiki/Explainable_artificial_intelligence

Explainability is critical for model developers and end-users

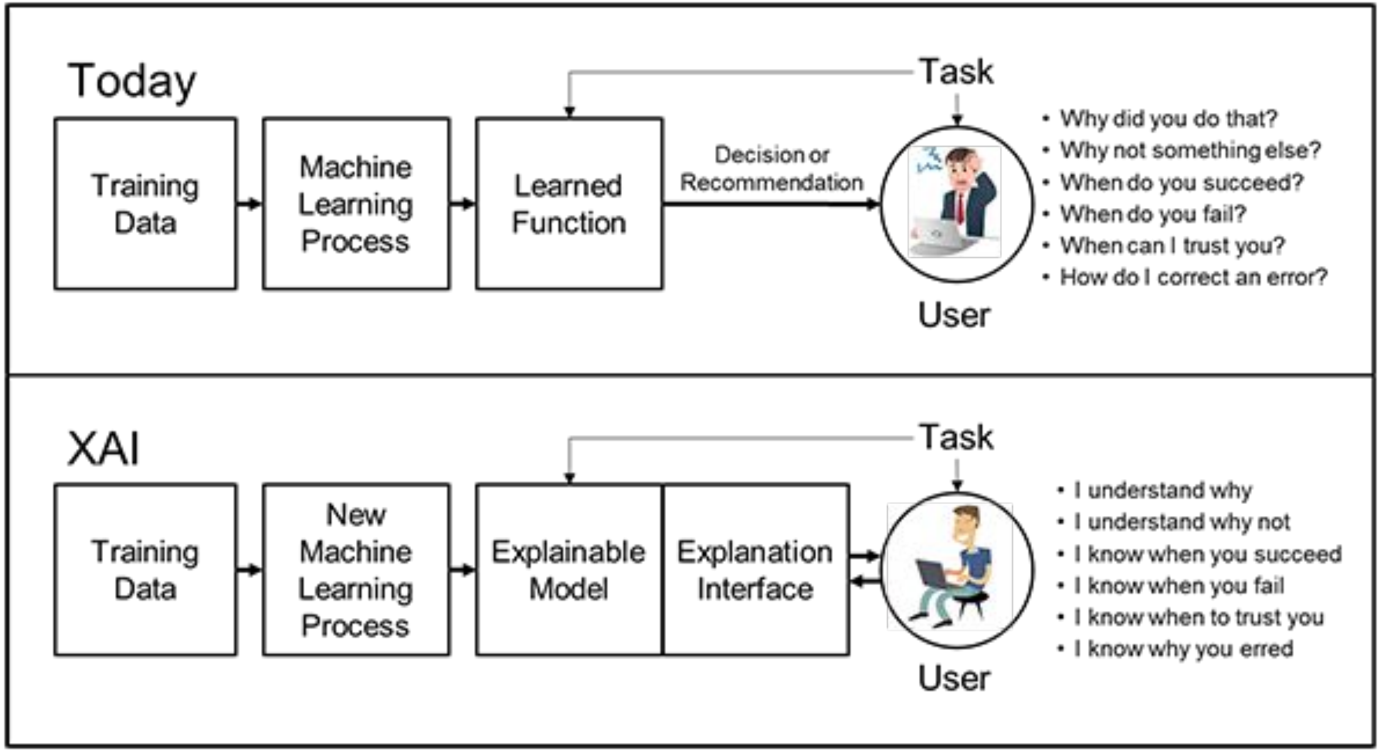
Respect

Assessing fairness of rules

Contesting and correcting decisions

Ability to change user behavior

https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=5569&context=flr


Today
Task
Training Data
Machine Learning Process
Learned Function
Decision or Recommendation
User
• Why did you do that?
• Why not something else?
• When do you succeed?
• When do you fail?
• When can I trust you?
• How do I correct an error?
XAI
Task
Training Data
New Machine Learning Process
Explainable Model
Explanation Interface
User
• I understand why
• I understand why not
• I know when you succeed
• I know when you fail
• I know when to trust you
• I know why you erred

Logic-based systems have strong explainability

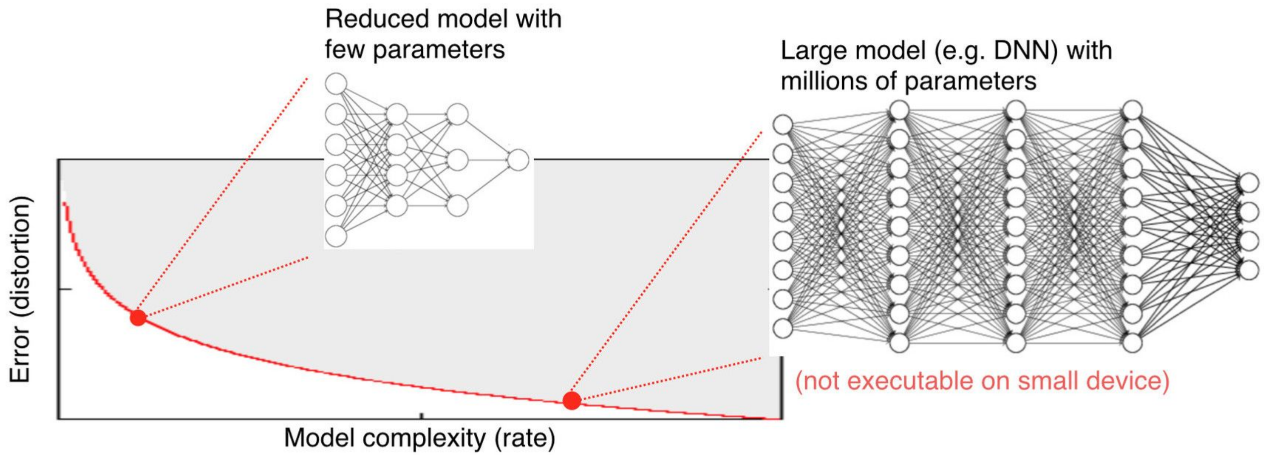
Transparent reasoning process (explicit knowledge representation, inference rules)

Justification of decision-making

Formal verification

More complex ML-based systems tend to be less explainable, but more performant

Neural networks with billions of parameters are more complex and inherently less explainable to humans



https://www.hhi.fraunhofer.de/en/departments/ai/research-groups/efficient-deep-learning/research-topics/neural-network-reduction-and-optimization.html

Interpretability

Understanding why a model generates certain outputs by understanding how the model's weights and features determine the given output.

"Can we understand how the model works internally by examining its structure, parameters, or learned representations?"

https://en.wikipedia.org/wiki/Explainable_artificial_intelligence
Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608.1

Interpretability is critical for model developers and end-users

Steerability

Identifying influential data features

Identifying influential representations

Verifiability

https://ir.lawnet.fordham.edu/cgi/viewcontent.cgi?article=5569&context=flr

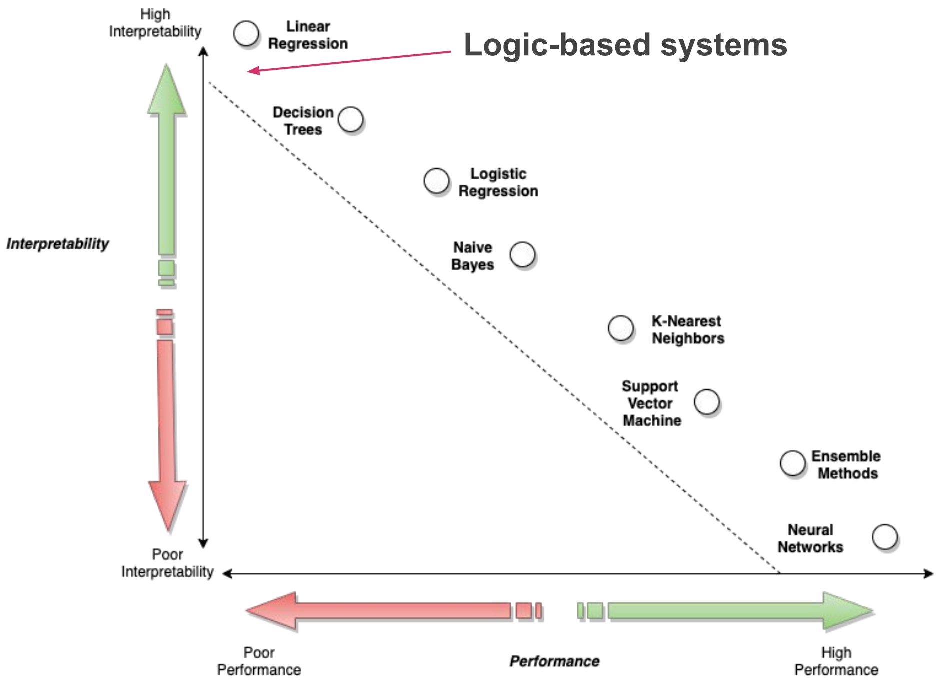
Logic-based systems have strong interpretability

Explicit and human-readable knowledge representation

Justification of decision-making

Modularity and formal semantics

More complex ML-based systems tend to be less interpretable, but more performant



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

Emerging best practices when designing explainable / interpretable AI systems

Provide clear documentation (Data and Model Cards)

Standardized docs outlining characteristics, limitations, and intended use

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model Cards for Model Reporting. In Proceedings of the Conference on Fairness, Accounta
Transparency (pp. 220-229).2

Emerging best practices when designing explainable / interpretable AI systems

Provide clear documentation (Data and Model Cards)

Standardized docs outlining characteristics, limitations, and intended use

Engage human stakeholders in evaluation

Design explainability and interpretability mechanisms that are understandable by end-users

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model Cards for Model Reporting. In Proceedings of the Conference on Fairness, Accounta
Transparency (pp. 220-229).2

Emerging best practices when designing explainable / interpretable AI systems

Provide clear documentation (Data and Model Cards)

Standardized docs outlining characteristics, limitations, and intended use

Engage human stakeholders in evaluation

Design explainability and interpretability mechanisms that are understandable by end-users

Consider when to use simpler vs. more complex models

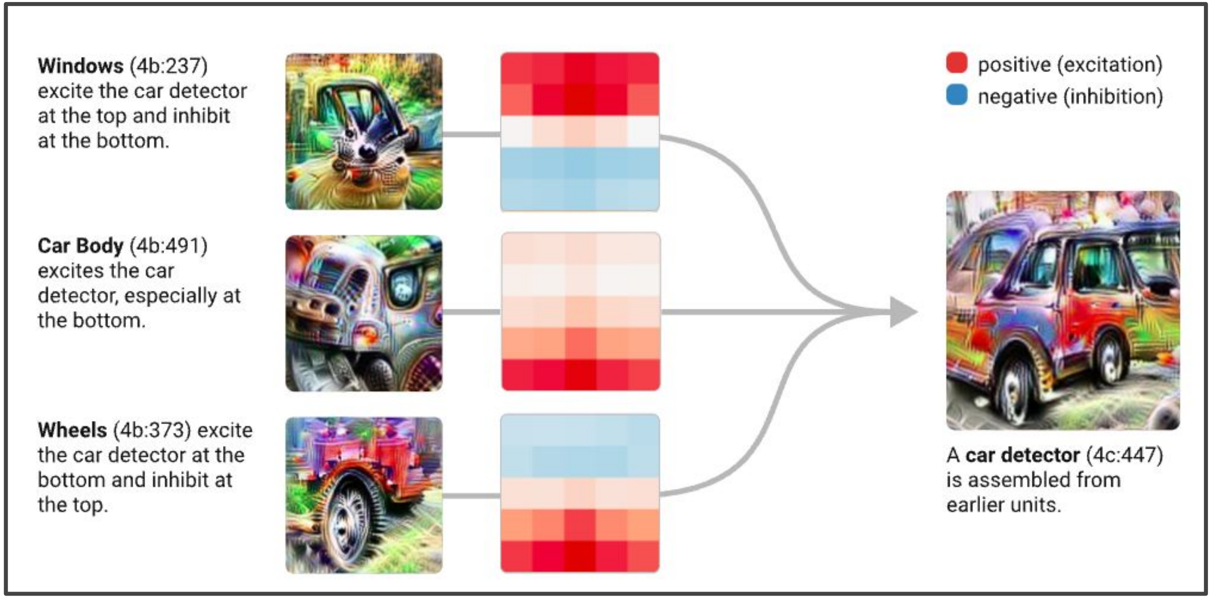
Simpler models (rule-based systems, or decision trees) are easier to understand

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019, January). Model Cards for Model Reporting. In Proceedings of the Conference on Fairness, Accounta
Transparency (pp. 220-229).2

Emerging research on explainability and interpretability

Mechanistic interpretability

Reverse engineer neural networks, similar to reverse engineering a compiled binary computer program, to understand model internals



https://www.lesswrong.com/posts/jLAvJt8wuSFySN975/mechanistic-interpretability-quickstart-guide

Emerging research on explainability and interpretability

Mechanistic interpretability

Reverse engineer neural networks, similar to reverse engineering a compiled binary computer program, to understand model internals

Local explanation techniques

Saliency maps or feature attributions are examples of trying to explain model output by understanding how input is used

https://christophm.github.io/interpretable-ml-book/pixel-attribution.html

Emerging research on explainability and interpretability

HATECHECK: Functional Tests for Hate Speech Detection Models

Paul Röttger[1,2], **Bertram Vidgen**[2], **Dong Nguyen**[3], **Zeerak Waseem**[4],
Helen Margetts[1,2], and **Janet B. Pierrehumbert**[1]

[1]University of Oxford
[2]The Alan Turing Institute
[3]Utrecht University
[4]University of Sheffield

internals

how input is used

Auditing methods

Top-down approach to understand how models behave on carefully constructed examples

Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. (2021). HateCheck: Functional Tests for Hate Speech Detection Models. In Proceedings of the 59th Annual Meetin
Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 41-58).[4]

Interpretability research supports general AI research

Work by Zhengxuan Wu et al. demonstrates that interpretability methods can be adapted to create an accurate and efficient fine-tuning mechanism (~10-50x more efficient than existing best methods)

ReFT: Representation Finetuning for Language Models

Zhengxuan Wu[*†] **Aryaman Arora**[*†] **Zheng Wang**[†] **Atticus Geiger**[‡]
Dan Jurafsky[†] **Christopher D. Manning**[†] **Christopher Potts**[†]
[†]Stanford University [‡]Pr(Ai)²R Group
{wuzhengx,aryamana,peterwz,atticusg,jurafsky,manning,cgpotts}@stanford.edu

Wu, Z., Arora, A., Wang, Z., Geiger, A., Jurafsky, D., Manning, C. D., & Potts, C. (2024). ReFT: Representation Finetuning for Language Models. arXiv preprint arXiv:2404.03592.[3]

Thank you!

Please reach out on Ed if you have any feedback.