

# Beyond the Black Box: Data Ethics, Explainable AI, and Human Oversight

2025 Joint Statistical Meeting, Nashville, TN

---

Chong Ho Alex Yu, Ph.D., D. Phil.

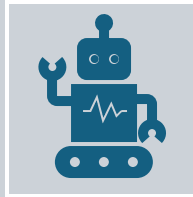




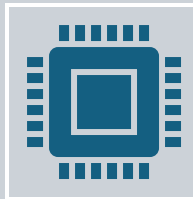
**“The black box society is one that invites a sort of automation of willful blindness.”**

- Frank Pasquale, Legal Scholar

### 3. Black box nature in almost all technologies



The "black box" nature is not unique to AI; many technologies are opaque to their users.



For example, most drivers do not understand how an engine works. Most computer users have no ideas how the CPU processes data.

## 4. Differences between AI and other technologies



**Scale and complexity:** AI systems operate on a scale of complexity far beyond most other technologies.



**Decision-making autonomy:** Unlike a car engine or computer hardware, AI systems could make autonomous decisions (e.g., self-drive car).



**Evolving nature:** AI systems can learn and change over time, potentially in ways not fully anticipated by their creators.



**AI hallucinations:** AI systems tend to make up answers!



## 5. Transparency

- Transparency refers to the openness and accessibility of information about an **AI system as whole**.
- It involves making the processes, data, algorithms, and decisions of an AI system available and understandable to stakeholders.





## 6. Transparency

- **Data Transparency:** Providing information about the data used to train the AI system, including sources, preprocessing methods, and potential biases.
- **Algorithmic Transparency:** Disclosing the algorithms and models used in the AI system, including their design, architecture, and functioning.
- **Transparency in Decision-Making Process:** Offering insights into how decisions are made by the AI system, including the logic and rules applied at each step.



## 7. Explainability

- Explainability refers to the ability to provide understandable and transparent explanations for how an AI system arrives at a particular decision or outcome.





## 8. Issues of explainability

- **Accountability:**
  - Without clear explanations, it becomes difficult to hold AI systems or their developers accountable for errors or biases. This is particularly problematic in high-stakes domains like healthcare, criminal justice, or finance, where decisions can significantly impact people's lives.



## 9. Issues of explainability

**TRUST:** USERS AND STAKEHOLDERS ARE LESS LIKELY TO TRUST AI SYSTEMS IF THEY CANNOT UNDERSTAND HOW DECISIONS ARE MADE.

**BIAS AND FAIRNESS:** IF THE DECISION-MAKING PROCESS IS NOT EXPLAINABLE, IT IS HARD TO IDENTIFY AND MITIGATE BIASES WITHIN THE AI SYSTEM.



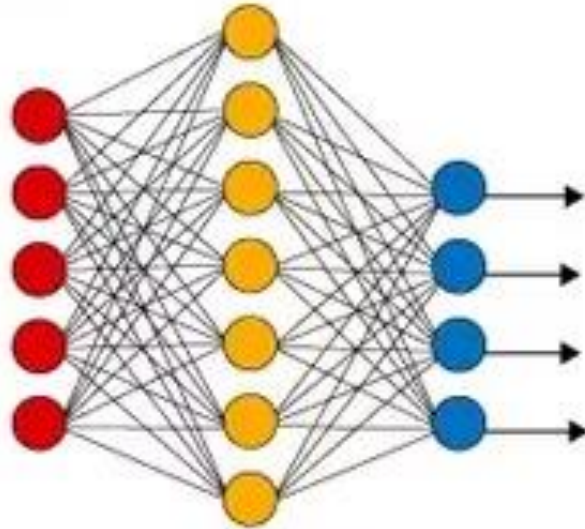
# 10. Interpretability

- **Interpretability** refers to the degree of simplicity to which a human can understand the cause of a decision made by an AI system.
- An interpretable model is one where the reasoning process is clear and understandable.



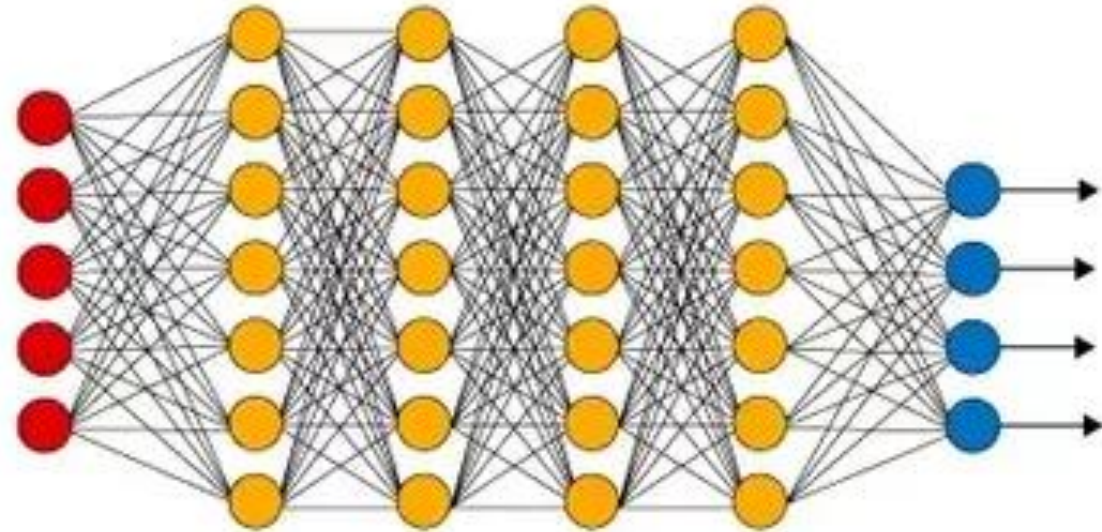


Simple Neural Network



● Input Layer

Deep Learning Neural Network



● Hidden Layer

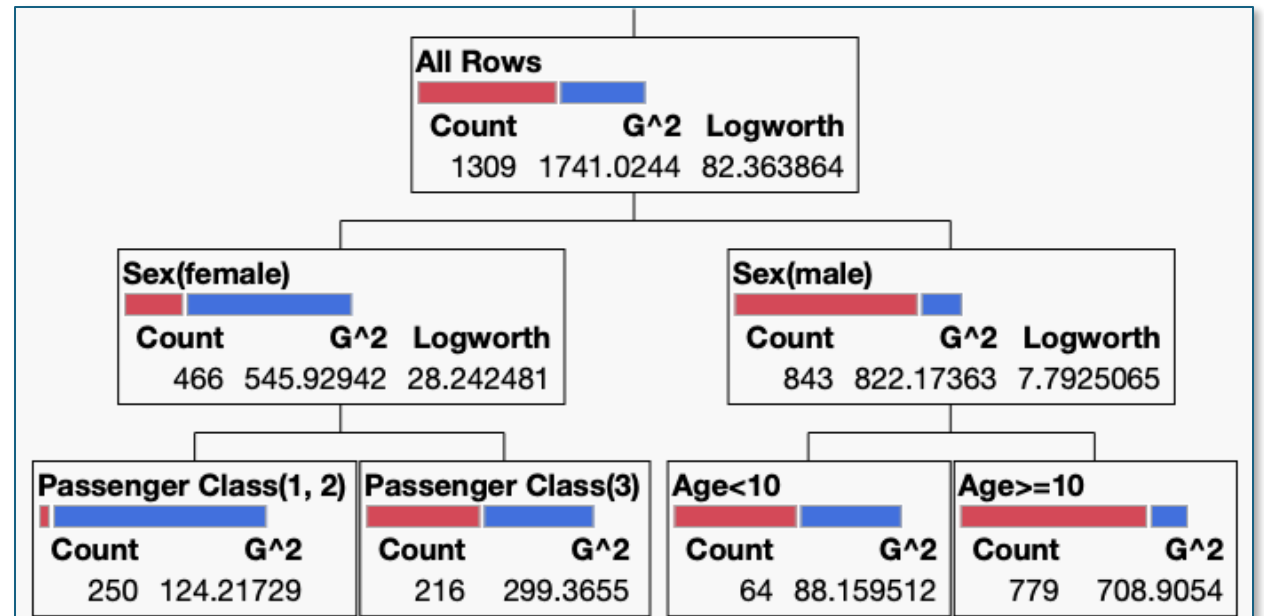
● Output Layer

## 11. Issues of interpretability

- **Complexity of Models:** Many modern AI models, such as deep neural networks, are inherently complex and involve numerous hidden layers and parameters.

# 12. Issues of interpretability

- **Transparency vs. Performance Trade-off:** Simpler models like decision trees or linear regression are more interpretable but might not perform as well as complex models like deep neural networks.







**GDPR**

EU General Data  
Protection Regulation  
**25 May 2018**

### 13. Issues of interpretability

- **Legal and Ethical Implications:** Regulations such as the European Union's General Data Protection Regulation (GDPR) include provisions for the "right to explanation," which requires that individuals can obtain an explanation for decisions made by automated systems.

## 14. Differences between transparency and explainability

### Scope:

- **Transparency:** Encompasses a broad view of the AI system, including data, algorithms, and processes.
- **Explainability:** Focuses specifically on making individual decisions or outcomes understandable.


### Purpose:

- **Transparency:** Aims to build trust by being open about how the system operates and what it is based upon.
- **Explainability:** Aims to make specific decisions or predictions understandable and interpretable by humans.

## 15. Difference between transparency and explainability

- **Approach:**
  - **Transparency:** Often involves documentation, open access to code and data, detailed descriptions of algorithms, and inputs from multiple stakeholders.
  - **Explainability:** Involves developing methods to articulate the decision-making process in a human-understandable manner, such as through visualizations, natural language explanations, or simplified models.





## 16. Difference between explainability and interpretability

- **Scope:**
  - **Explainability:** Broader concept that includes methods to provide understandable explanations for complex, black-box models, such as deep learning and ensemble methods.
  - **Interpretability:** Narrowly Focuses on the intrinsic understandability of the model itself, often favoring simpler, transparent models. Typical examples include decision trees and support vector machines.



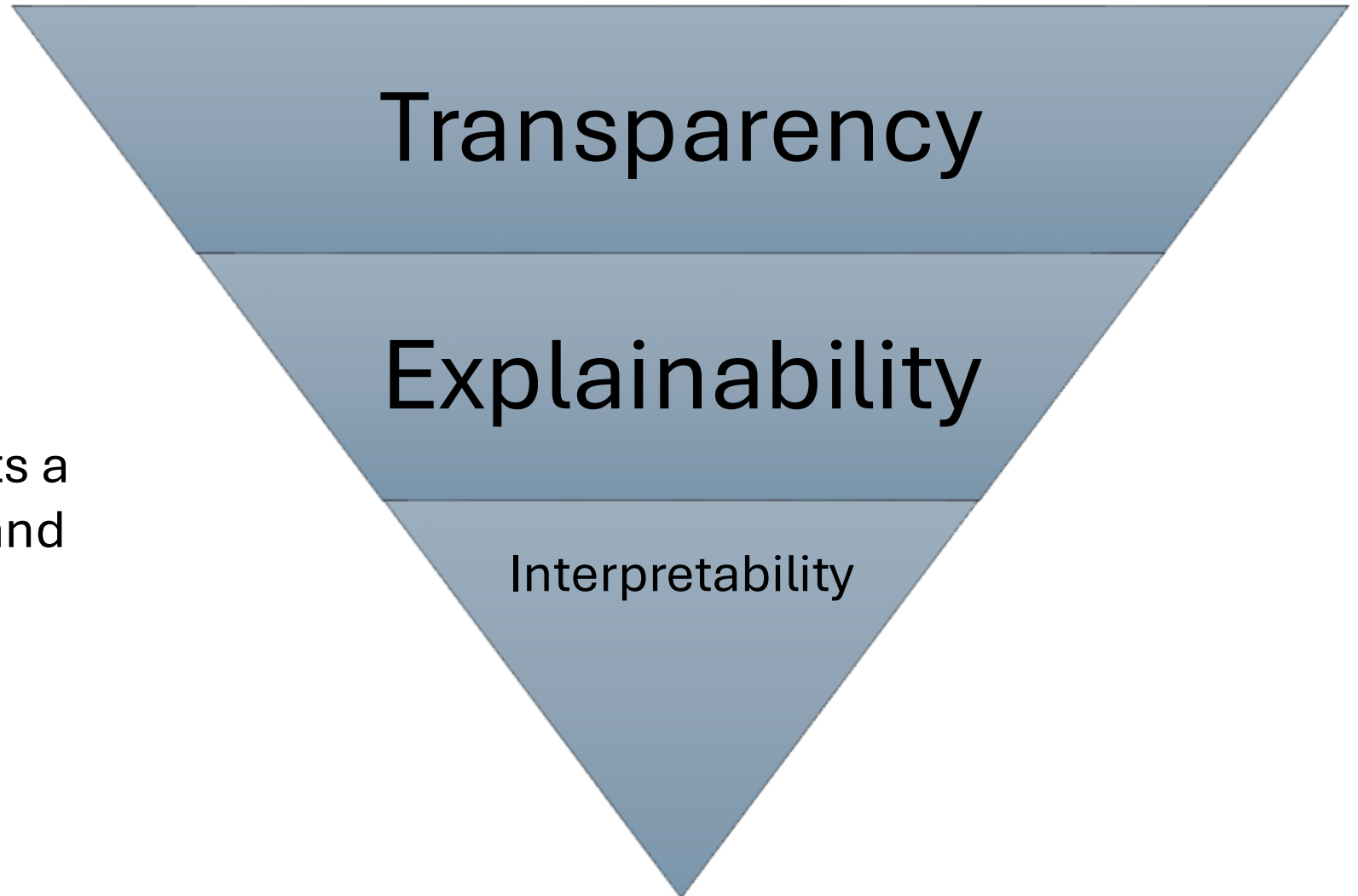
## 17. Difference between explainability and interpretability

- **Approach:**
  - **Explainability:** Involves creating post-hoc explanations to make complex models' decisions understandable. Typical examples include SHAP values, LIME, and data visualizations.
  - **Interpretability:** Typically involves simpler models that are directly understandable without additional explanatory tools.



# 18. Reversed triangle

- We can visualize transparency, explainability, and interpretability as a reversed triangle where each concept represents a different level of detail and focus within the AI and data science system.



# 19. Concentric circles

- We can also picture their relationships as concentric circles.
  - **Transparency (Outer ring):** Covers the entire system, including data collection, model training, deployment, and decision-making processes.
  - **Explainability (middle):** Centers on making the model's behavior and decision processes understandable.
  - **Interpretability (inside):** Focuses on using simpler models or techniques that are inherently understandable by humans.



## 20. COMPAS (2016)

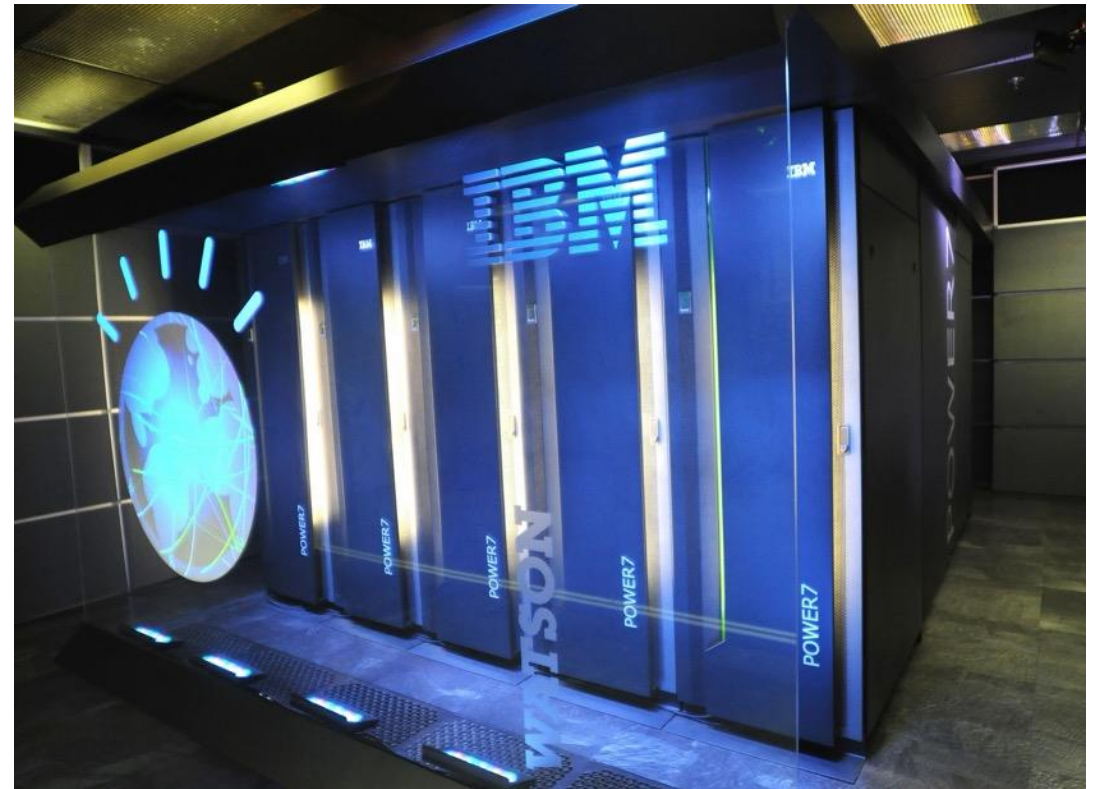
- The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm, used in the U.S. criminal justice system to predict the likelihood of a defendant reoffending, faced significant scrutiny. A ProPublica investigation revealed that the algorithm was biased against Black defendants, predicting them to be at higher risk of reoffending compared to white defendants.
- The lack of transparency and explainability in how COMPAS made its predictions raised ethical concerns about fairness and accountability in AI systems.



# 21. IBM Watson Health (2018)

IBM's Watson for Oncology project faced criticism for making "unsafe and incorrect" cancer treatment recommendations.

The lack of transparency in how Watson reached its conclusions raised concerns about its reliability in healthcare settings.



## 22. Facebook Ad Targeting (2019)

The Facebook logo, consisting of the word "facebook" in white lowercase letters on a blue background. The background of the logo is a solid blue shape with a torn, irregular edge on the right side.

- Facebook involved in multiple controversies regarding its ad targeting algorithms. In 2019, the U.S. Department of Housing and Urban Development (HUD) charged Facebook with enabling discriminatory housing advertisements.
- The AI algorithms allowed advertisers to exclude certain demographics from seeing ads, which is illegal under the Fair Housing Act.
- This case indicates the importance of transparency and explainability to ensure compliance with anti-discrimination laws.

## 23. Apple Card Gender Bias (2019)



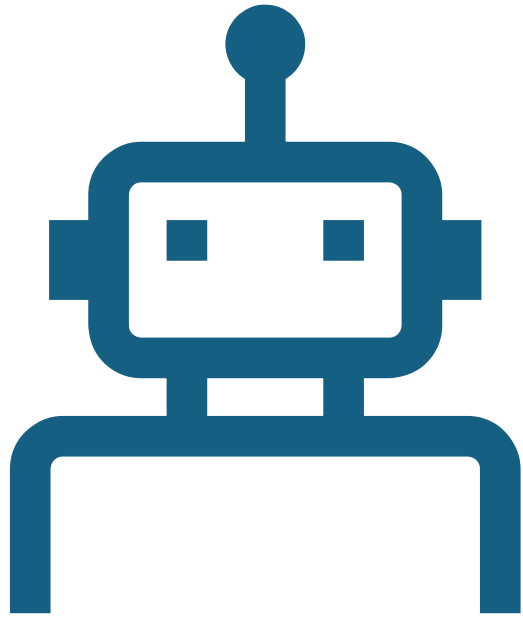
- In 2019, Apple and Goldman Sachs faced a scandal when users reported that the Apple Card's credit limit algorithm appeared to discriminate against women.
- Despite similar financial profiles, women were given significantly lower credit limits than men.
- Although it may not be intentional, the opacity of the algorithm's decision-making process made it difficult to understand and address the differential outcomes.



## 24. UK A-Level Grading Algorithm (2020)

- +
- 
- 
- During the COVID-19 pandemic, the UK used an algorithm to determine students' A-level grades.
- The algorithm's lack of transparency and apparent bias against students from disadvantaged backgrounds, resulting in widespread protests and eventual abandonment of the system.





## 25. Explainable AI

- Aims to provide insights into the reasoning behind AI decisions, countering the "black-box".
  - **Data Explainability:** Understanding the data used to train AI models.
  - **Model Explainability:** Designing models that are inherently interpretable.
  - **Post-hoc Explainability:** Providing explanations after the model has made a decision, often through techniques like feature importance and visualization.

## 26. Audit trail

- An audit trail in research refers to the systematic documentation of the research process, which provides a clear, transparent record of how the research was conducted.
- In the context of AI, an audit trail can enhance transparency, accountability, and trustworthiness.



## 27. Hybrid Models

- Combining interpretable models with complex models to achieve a balance between performance and interpretability. For example, using a simple model to provide explanations for a more complex model's decisions.



## 28. Examples of Hybrid models

- **Healthcare:**
  - **Disease Diagnosis:** In medical diagnostics, hybrid models can combine the interpretability of decision trees with the accuracy of neural networks.
  - For example, a model might use a decision tree to identify key symptoms and patient history, and then use a neural network to analyze medical images. This approach allows doctors to understand the diagnostic process while leveraging the predictive power of deep learning.



# 29. Why keeping humans in the loop

- Keeping **humans in the loop** should be considered the most important solution to the "black box" nature of AI.
- Human oversight ensures there's a responsible party who can be held accountable for decisions made by AI systems. This is crucial in high-stakes domains like healthcare, finance, and law enforcement.
- AI systems can struggle with unusual or rare situations. Human oversight allows for intervention in these edge cases where the AI's decision might be inappropriate or dangerous.

## 30. For more information

- [chonghoyu@gmail.com](mailto:chonghoyu@gmail.com)
- My YouTube channel on AI and data analytics:  
<https://www.youtube.com/@datafrontiers>



- My blog on data science and machine learning:  
<https://creative-wisdom.me/blog>

