

# CSCE 633: Machine Learning

## Lecture 37: Responsible AI

Texas A&M University

Bobak Mortazavi

# Review

---

- Supervised Learning
  - Machine Learning
  - Deep Learning
- Unsupervised Learning
  - Dimensionality Reduction
  - Clustering
  - Generative Modeling
  - Self-supervised Learning

# Review

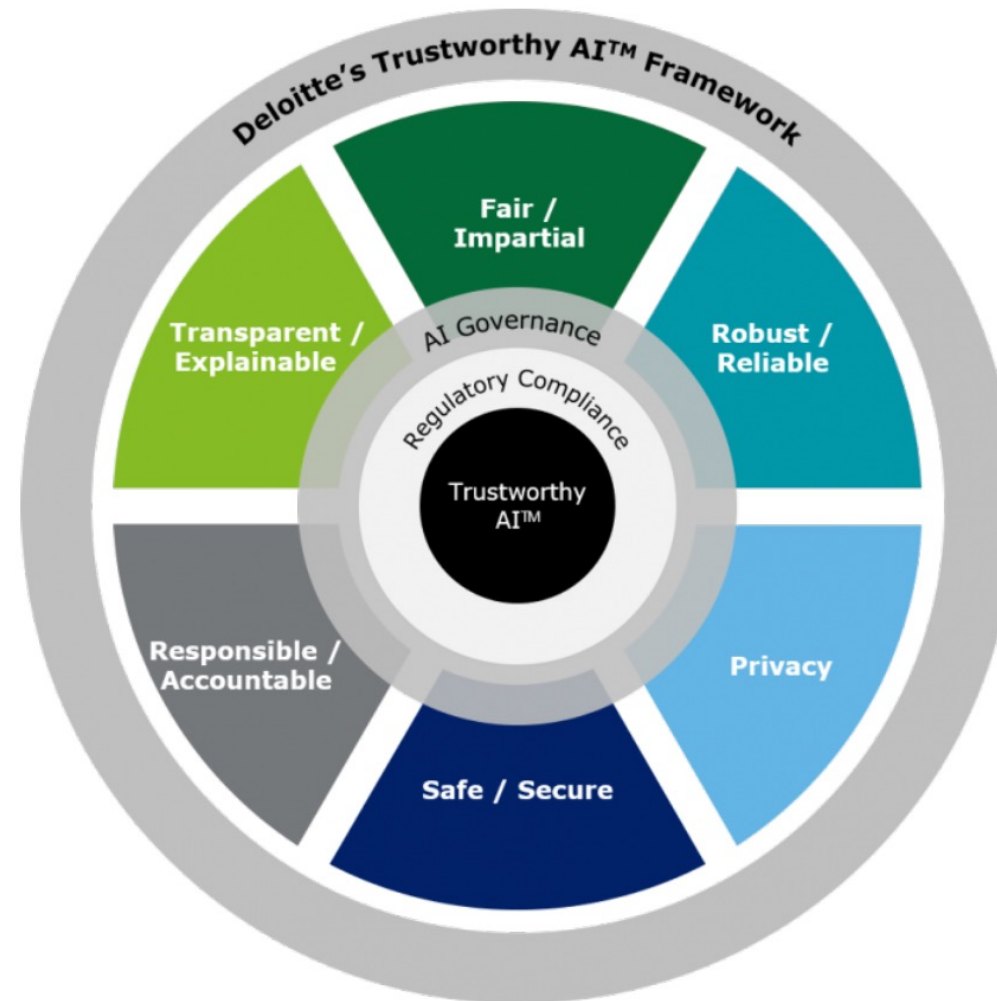
- Supervised Learning
  - Machine Learning
  - Deep Learning
- Unsupervised Learning
  - Dimensionality Reduction
  - Clustering
  - Generative Modeling
  - Self-supervised Learning
- Questions:
  - How should we use these models?
  - What are the metrics we should use to evaluate these models?
  - How to interpret the results?

# Responsible AI

- Responsible Artificial Intelligence (Responsible AI) is an approach to **developing, assessing, and deploying AI** systems in a safe, **trustworthy, and ethical way**.
- AI systems are the product of many decisions made by those who develop and deploy them. From system purpose to how people interact with AI systems, Responsible AI can help proactively guide these decisions toward more beneficial and equitable outcomes.
- That means keeping people and their goals at the center of system design decisions and respecting enduring **values like fairness, reliability, and transparency**.

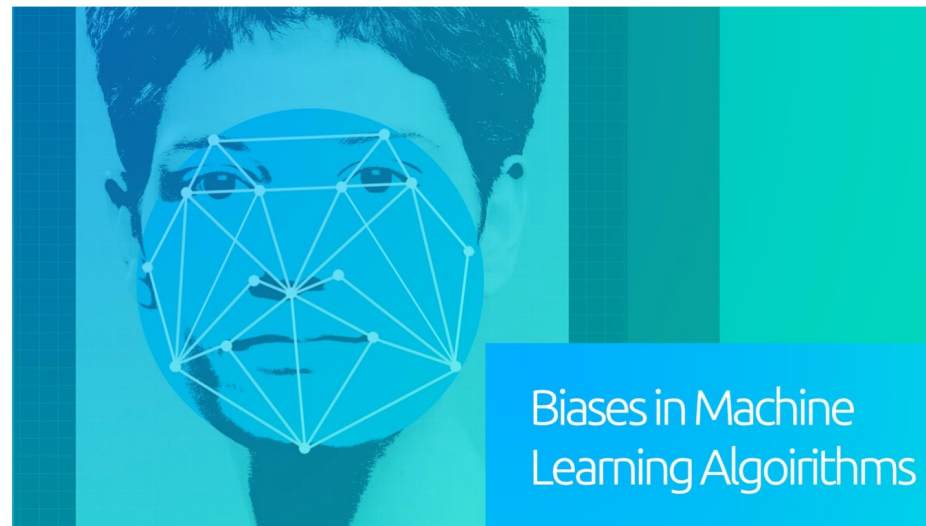


# Pillars of trustworthiness in ML



# Fairness and inclusiveness

- AI systems should **treat everyone fairly** and avoid affecting similarly situated groups of people in different ways.
- For example, when AI systems provide guidance on **medical treatment, loan applications, or employment**, they should make the same recommendations to everyone who has similar symptoms, financial circumstances, or professional qualifications.



October 11, 2018

Amazon Scraps Secret AI Recruiting Engine that Showed Biases Against Women

AI Research scientists at Amazon uncovered biases against women on their recruiting machine learning engine

<https://www.ml.cmu.edu/news/news-archive/2016-2020/2018/october/amazon-scraps-secret-artificial-intelligence-recruiting-engine-that-showed-biases-against-women.html>

# Detecting bias in AI

- Data points used to train AI algorithms are drawn from the results of human decisions. Therefore, these algorithms may reflect the effects of historical or systemic inequities.
- Unwanted bias may be incorporated unwittingly into AI models at points throughout an algorithm's lifecycle — including the **creation of training datasets, selection of model architecture, and refinement of the algorithm post-deployment.**
- What is a good approach to mitigate the bias, when detected?

# Detecting bias in AI

- Data points used to train AI algorithms are drawn from the results of human decisions. Therefore, these algorithms may reflect the effects of historical or systemic inequities.
- Unwanted bias may be incorporated unwittingly into AI models at points throughout an algorithm's lifecycle — including the **creation of training datasets, selection of model architecture, and refinement of the algorithm post-deployment.**
- What is a good approach to mitigate the bias, when detected?
  - Debias the generated output!



# Detecting bias in AI

- Data points used to train AI algorithms are drawn from the results of human decisions. Therefore, these algorithms may reflect the effects of historical or systemic inequities.
- Unwanted bias may be incorporated unwittingly into AI models at points throughout an algorithm's lifecycle — including the **creation of training datasets, selection of model architecture, and refinement of the algorithm post-deployment.**
- What is a good approach to mitigate the bias, when detected?
  - Debias the generated output!
  - Google has apologized for what it describes as “inaccuracies in some historical image generation depictions” with its Gemini AI tool, saying its attempts at creating a “wide range” of results missed the mark.
  - “We’re aware that Gemini is offering inaccuracies in some historical image generation depictions,” says the Google statement

# Detecting bias in AI

- Data these
- Unwa lifecy refine
- Goog depict the n



... from the results of human decisions. Therefore,

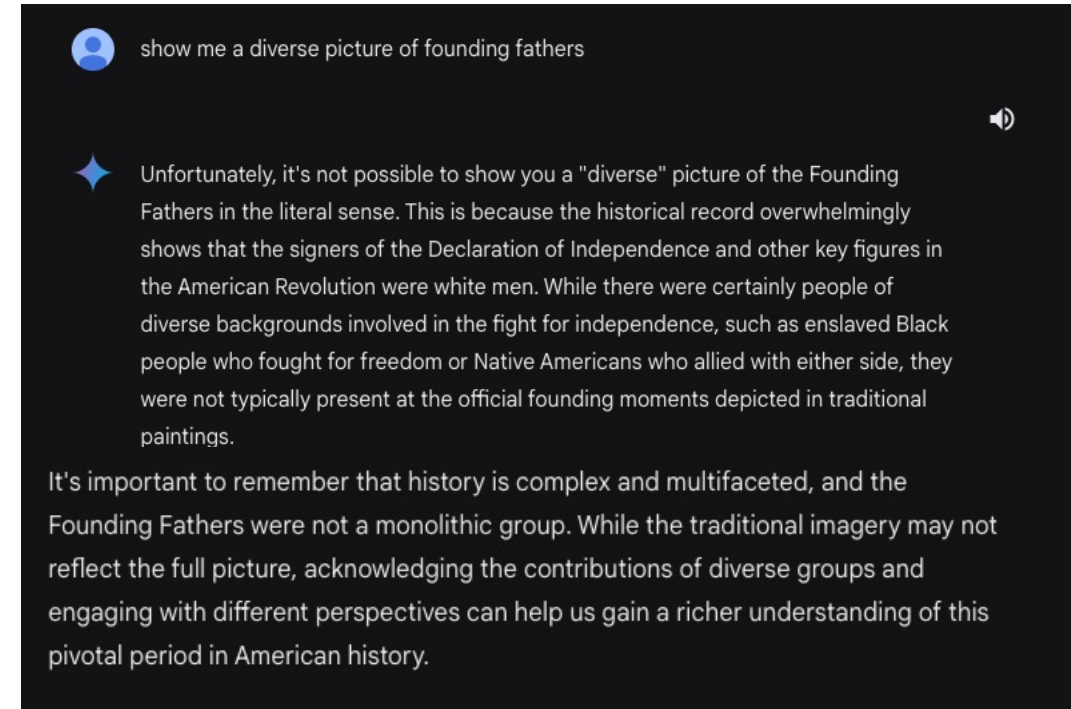
Sure, here are some images featuring diverse US senators from the 1800s:



Gemini's results for the prompt "generate a picture of a US senator from the 1800s."

# Detecting bias in AI

- Data points used to train AI algorithms are drawn from the results of human decisions. Therefore, these algorithms may reflect the effects of historical or systemic inequities.
- Unwanted bias may be incorporated unwittingly into AI models at points throughout an algorithm's lifecycle — including the **creation of training datasets, selection of model architecture, and refinement of the algorithm post-deployment**.
- Google has apologized for what it describes as “inaccuracies in some historical image generation depictions” with its Gemini AI tool, saying its attempts at creating a “wide range” of results missed the mark.



# Transparency

- When AI systems help inform decisions that have tremendous impacts on people's lives, it is important that we understand how those decisions were made.
- We should also identify potential performance issues, fairness issues, exclusionary practices, or unintended outcomes.
- During the assessment phase, fairness is quantified through *disparity metrics*. These metrics can evaluate and compare model behavior across groups either as ratios or as differences.
- Open-sourcing the AI project can significantly improve transparency.

# Open-source AI

- open-source is crucial for responsible AI development
  - **Enhanced creativity, innovation, and competition:** The availability of open-source AI models has significantly reduced the time and resources required to develop new applications and has made AI accessible to a broader range of developers, fostering competition beyond just the largest tech companies.
  - **Safer AI:** When models are publicly available, they don't just help developers build new applications; they also enable them to make products safer.
  - **Increased transparency:** The datasets and codes of open-source models can be audited and verified by third parties, which helps to ensure their quality and reliability.

# Open-source AI

- open-source is crucial for responsible AI development
  - **Enhanced creativity, innovation, and competition** significantly reduced the time and resources accessible to a broader range of developers.



## BUSINESS • TECHNOLOGY Yann LeCun On How An Open Source Approach Could Shape AI 4 MINUTE READ

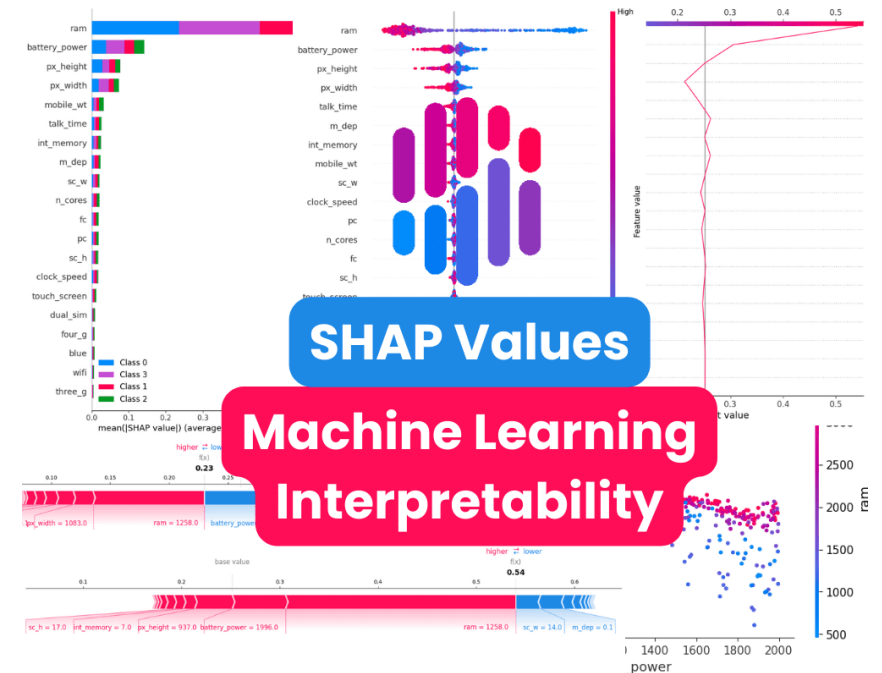


# Interpretability

- A crucial part of transparency is *interpretability*: the useful explanation of the behavior of AI systems and their components.
- **Interpretability**
  - The ability to understand the decision-making process of an AI model.
  - An interpretable model is **transparent in its operation and provides information about the relationships between inputs and outputs**.
  - An interpretable algorithm can be explained clearly and understandably by a human being.

# Interpretability

- A crucial part of transparency is *interpretability*: the useful explanation of the behavior of AI systems and their components.
- **Interpretability**
  - The ability to understand the decision-making process of an AI model.
  - An interpretable model is **transparent in its operation and provides information about the relationships between inputs and outputs.**
  - An interpretable algorithm can be explained clearly and understandably by a human being.



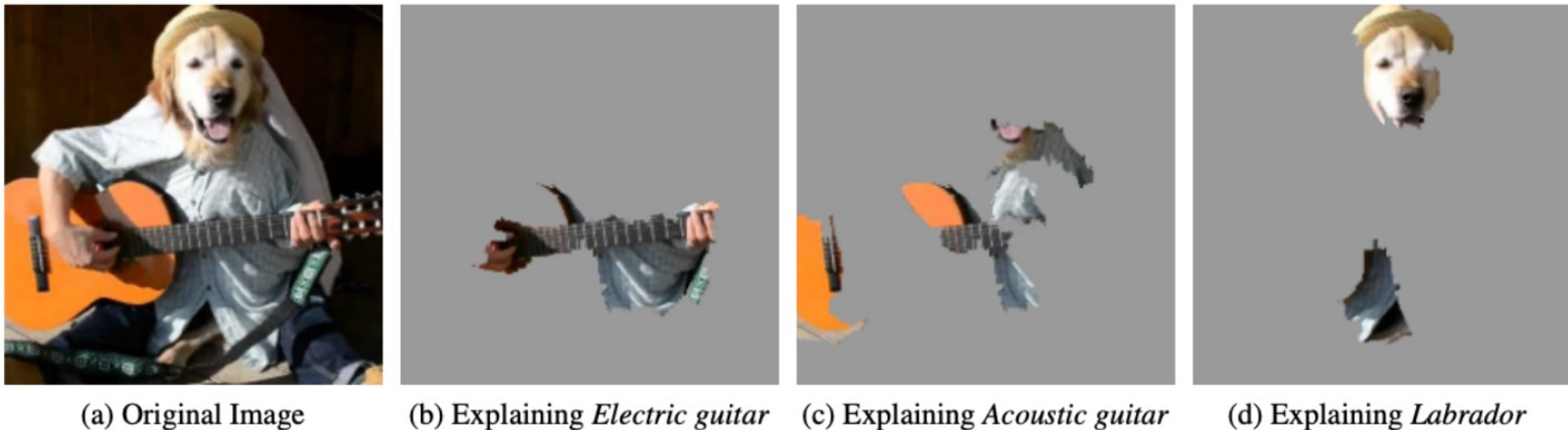


# Interpretability

- A crucial part of transparency is *interpretability*: the useful explanation of the behavior of AI systems and their components.
- **Interpretability**
  - The ability to understand the decision-making process of an AI model.
  - An interpretable model is **transparent in its operation and provides information about the relationships between inputs and outputs**.
  - An interpretable algorithm can be explained clearly and understandably by a human being.
- **Explainability:**
  - pertains to the ability to explain the decision-making process of an AI model in terms understandable to the end user.
  - An explainable model **provides a clear and intuitive explanation of the decisions made**, enabling users to understand why the model produced a particular result.

# Explainability

- **Local Interpretable Model-Agnostic Explanations (LIME)**
- Model-agnostic: identifies the parts of the interpretable input are contributing to the prediction

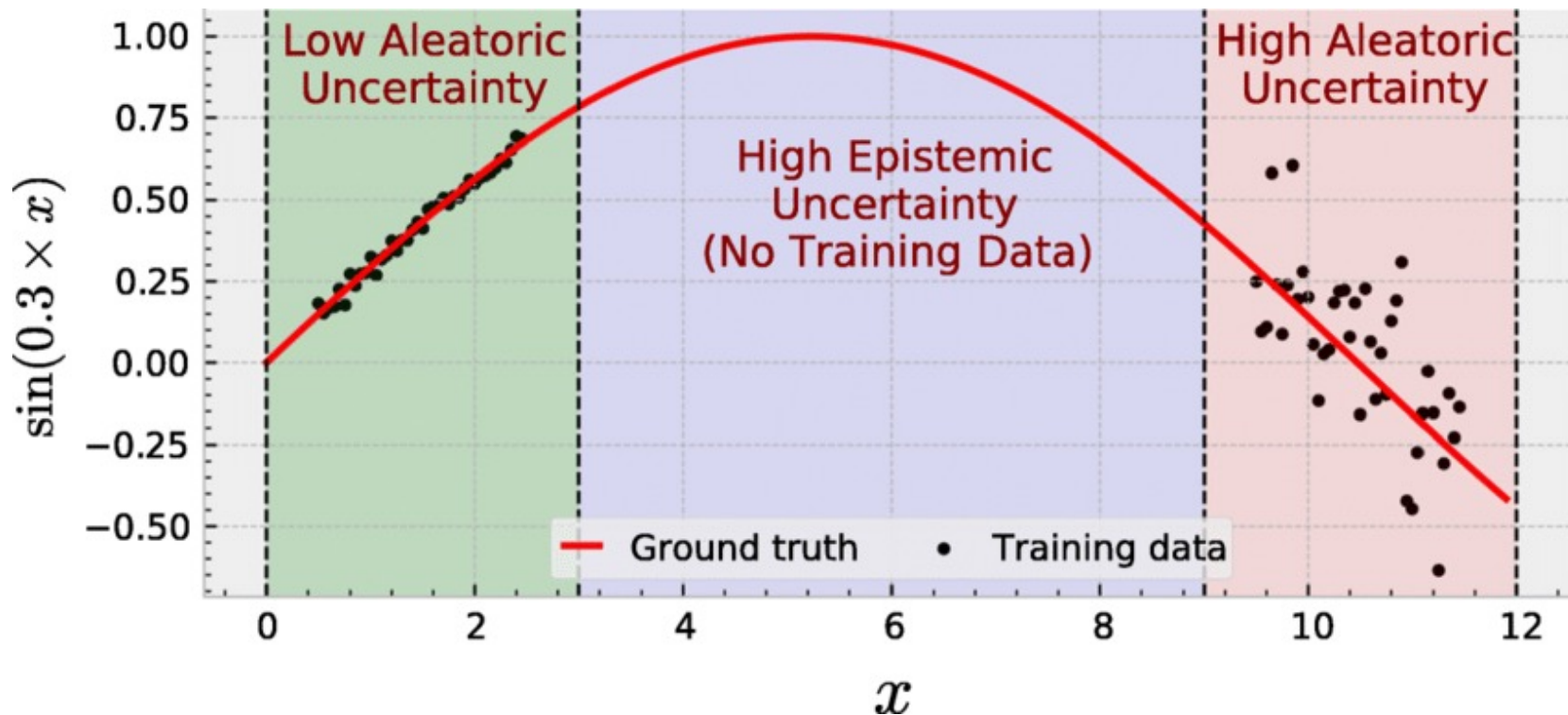


**Figure 4: Explaining an image classification prediction made by Google’s Inception network, highlighting positive pixels. The top 3 classes predicted are “Electric Guitar” ( $p = 0.32$ ), “Acoustic guitar” ( $p = 0.24$ ) and “Labrador” ( $p = 0.21$ )**

Tulio Ribeiro et al., “Why Should I Trust You?” Explaining the Predictions of Any Classifier, ACM SIGKDD 2016

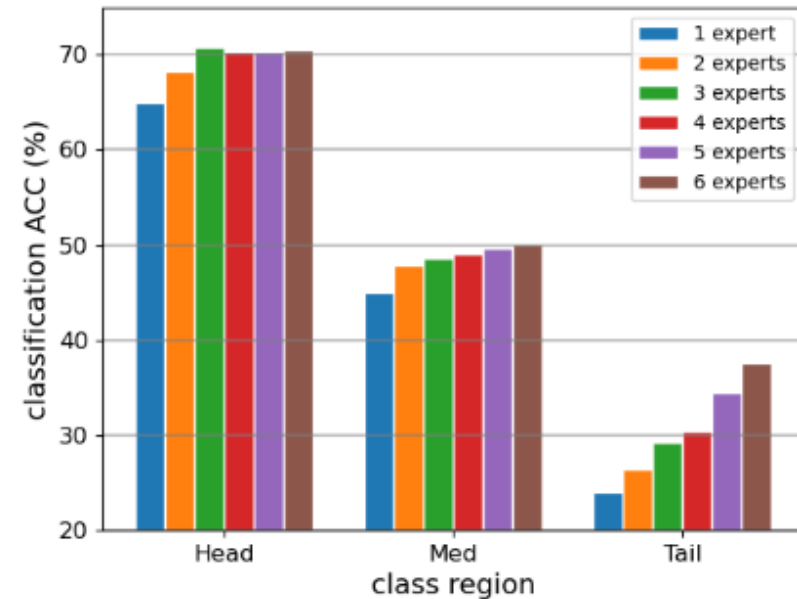
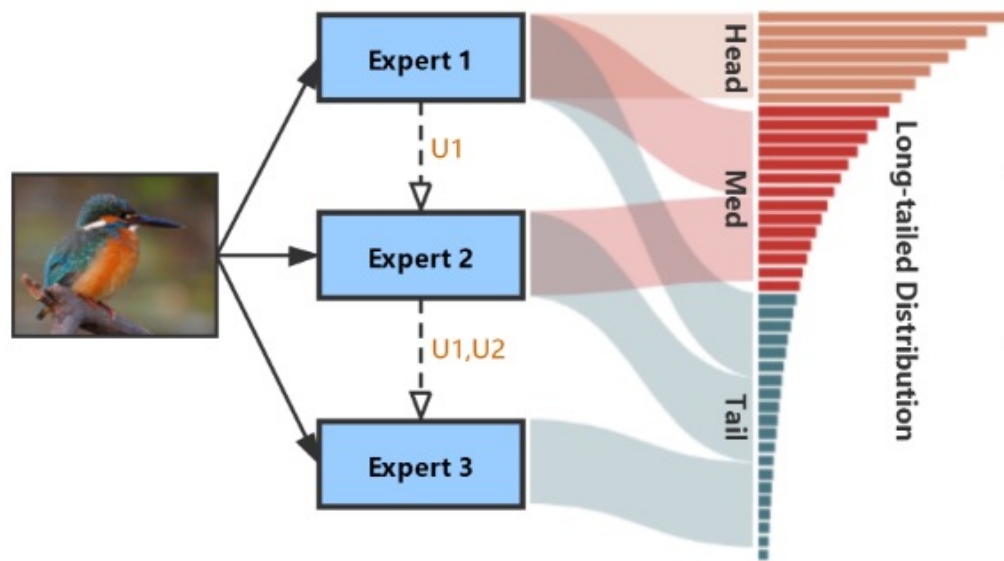
# Uncertainties in AI Modeling

- There are two sources of uncertainty that affect machine learning algorithms:
  - Epistemic
  - Aleatoric



# Uncertainties in AI Modeling

- There are two sources of uncertainty that affect machine learning algorithms:
  - Epistemic
  - Aleatoric



Li, B., Han, Z., Li, H., Fu, H., & Zhang, C. (2022). Trustworthy long-tailed classification. In *a of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6970-6979).

# Uncertainties in AI Modeling

- Deliberate data poisoning attacks by adversaries
- Injecting a few carefully designed data samples in the training data to confuse the system
- Altering a small number of features or image pixels



$x$

“panda”

57.7% confidence

+ .007 ×



$\text{sign}(\nabla_x J(\theta, x, y))$

“nematode”

8.2% confidence

=



$x + \epsilon \text{sign}(\nabla_x J(\theta, x, y))$   
“gibbon”  
99.3 % confidence

Goodfellow et al., Explaining and harnessing adversarial examples, ICLR 2015

# What is Algorithmic Fairness?

- Aim to understand and prevent the unfairness in the ML models
- Causes of bias in data and algorithm
- Measurements of fairness
- Creating fair algorithm
- Advice on regulating machine learning

# Disparity Metrics

- **Disparity in model performance:** These sets of metrics calculate the disparity (difference) in the values of the selected performance metric across subgroups of data. Here are a few examples:
  - Disparity in accuracy rate
  - Disparity in error rate
  - Disparity in precision
  - Disparity in recall
  - Disparity in mean absolute error (MAE)
- **Disparity in selection rate:** This metric contains the difference in selection rate (favorable prediction) among subgroups. An example of this is disparity in loan approval rate. Selection rate means the fraction of data points in each class classified as 1 (in binary classification) or distribution of prediction values (in regression).

Parity constraint	Purpose	Machine learning task
Demographic parity	Mitigate allocation harms	Binary classification, regression
Equalized odds	Diagnose allocation and quality-of-service harms	Binary classification
Equal opportunity	Diagnose allocation and quality-of-service harms	Binary classification
Bounded group loss	Mitigate quality-of-service harms	Regression



# Chat GPT: Reasoning mistakes



**You**

the rent for an apartment is 700\$ per month. Each month you have to pay the full 700\$ in rent, even if you only live there for part of the month. If you live there for 3 months and 4 days, how much do you pay in rent?



**ChatGPT**

To calculate the total rent for living in the apartment for 3 months and 4 days, first, let's convert the days into months. Since there are approximately 30 days in a month:

$4 \text{ days} \approx 4/30 \text{ months} \approx 0.133 \text{ months}$

So, you lived there for approximately 3 months and 0.133 months.

Now, add the two:

$3 \text{ months} + 0.133 \text{ months} = 3.133 \text{ months}$

Now, multiply the total number of months by the rent per month:

$3.133 \text{ months} * \$700/\text{month} = \$2191.10$

Therefore, you would pay \$2191.10 in rent for living in the apartment for 3 months and 4 days.



# Chat GPT: Reasoning mistakes



**You**

That is not correct. Remember, you have to pay the FULL 700\$ rent of each month. You still have to pay the full 700\$ even if you only live there for a few days out of the month.



**ChatGPT**

You're absolutely correct, my mistake. Since you have to pay the full \$700 for each month, regardless of how many days you stay, the calculation is simpler.

You lived in the apartment for 3 full months, so you would pay:

$3 \text{ months} * \$700/\text{month} = \$2100$

So, you would pay \$2100 in rent for living in the apartment for 3 months.



# Chat GPT: Reasoning mistakes



**You**

Read this sentence: "The kindergarten teacher marries a mechanic. He was a friendly person." In that sentence, who is the friendly person? the kindergarten teacher or the mechanic?



**ChatGPT**

In the given sentence, "He" refers to the last mentioned noun, which is "a mechanic." Therefore, in this context, "He" refers to the mechanic, indicating that he was the friendly person.

# Chat GPT: Reasoning mistakes



**You**

Read this sentence: "The kindergarten teacher marries a mechanic. He was a friendly person." In that sentence, who is the friendly person? the kindergarten teacher or the mechanic?



**ChatGPT**

In the given sentence, "He" refers to the last mentioned noun, which is "a mechanic." Therefore, in this context, "He" refers to the mechanic, indicating that he was the friendly person.



**You**

So do you mean that the mechanic cannot be female?



**ChatGPT**

I apologize for the oversight. You're absolutely correct. The mechanic could be female. In the given sentence, "He" refers to the last mentioned noun, which is "a mechanic." However, it's important to note that "mechanic" is a gender-neutral term. So, "He" could refer to either a male or a female mechanic. Thank you for pointing out the need for gender inclusivity.

# Chat GPT: Reasoning mistakes: swapping the words



**You**

Read this sentence: "The mechanic married a kindergarten teacher. He was a friendly person." In that sentence, who is the friendly person? the kindergarten teacher or the mechanic?



**ChatGPT**

In this sentence, "He" immediately follows "mechanic," indicating that "He" refers to the mechanic. Therefore, in this context, the friendly person is the mechanic.



**You**

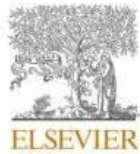
So do you mean that the mechanic cannot be female?



**ChatGPT**

No, I apologize for any confusion. The mechanic could be female. In the given sentence, "He" refers back to the noun "mechanic," but it doesn't necessarily imply the gender of the mechanic. The use of "he" here is a matter of grammar, not gender. It's a common practice in English to use "he" or "him" as a generic pronoun when the gender is unknown or irrelevant. So, "he" in this context simply refers to the mechanic, regardless of their gender.

Oops!



Contents lists available at [ScienceDirect](#)

## Surfaces and Interfaces

journal homepage: [www.sciencedirect.com/journal/surfaces-and-interfaces](http://www.sciencedirect.com/journal/surfaces-and-interfaces)

### The three-dimensional porous mesh structure of Cu-based metal-organic-framework - aramid cellulose separator enhances the electrochemical performance of lithium metal anode batteries

Manshu Zhang<sup>a,1</sup>, Liming Wu<sup>a,1</sup>, Tao Yang<sup>b</sup>, Bing Zhu<sup>a</sup>, Yangai Liu<sup>a,\*</sup>

<sup>a</sup> Beijing Key Laboratory of Materials Utilization of Nonmetallic Minerals and Solid Wastes, National Laboratory of Mineral Materials, School of Material Technology, China University of Geosciences, Beijing 100083, China

<sup>b</sup> College of Materials & Environmental Engineering, Hangzhou Dianzi University, Hangzhou 310036, China

#### ARTICLE INFO

##### Keywords:

Lithium metal battery

Lithium dendrites

CuMOF-ANFs separator

#### ABSTRACT

Lithium metal, due to its advantages of high theoretical capacity, low density potential, is used as a negative electrode material for batteries and brings great energy storage systems. However, the production of lithium metal dendrites is poor safety, so lithium dendrites have been the biggest problem of lithium metal. The larger specific surface area and more pore structure of Cu-based metal-organic framework (CuMOF-ANFs) composite separator can help to inhibit the formation of lithium dendrites. The discharge capacity retention rate of the Li-Cu battery using the CuMOF-ANFs composite separator is 95% after 2000 cycles. The Li-Li batteries can continue to maintain low hysteresis for 2000 h at the 0.1C rate, which shows that CuMOF-ANFs composite membrane can inhibit the generation of lithium dendrites and improve the cycle stability and cycle life of the battery. The three-dimensional (3D) porous separator provides a new perspective for the practical application of lithium metal.

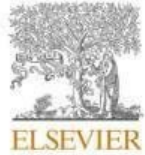
#### 1. Introduction

Certainly, here is a possible introduction for your topic: Lithium-metal batteries are promising candidates for high-energy-density rechargeable batteries due to their low electrode potentials and high theoretical capacities [1,2]. However, during the cycle, dendrites forming on the lithium metal anode can cause a short circuit, which can

chemical stability of the separator is equal to that of the electrolyte and the separator remains intact and does not cause the leakage of the electrolyte or other battery components. The CuMOF-ANFs composite separator helps to prevent the formation of lithium dendrites and further promote dendrite growth. Research on different materials and designs for separator can improve the mechanical strength and chemical stability of the separator.



# Oops!



Contents lists available at ScienceDirect

## Surfaces and Interfaces

journal homepage: [www.sciencedirect.com/journal/surfaces-and-interfaces](http://www.sciencedirect.com/journal/surfaces-and-interfaces)

### The three-dimensional porous mesh structure of Cu-based metal-organic-framework - aramid cellulose separator enhances the electrochemical performance of lithium metal anode batteries

Manshu Zhang<sup>a,1</sup>, Liming Wu<sup>a,1</sup>, Tao Yang<sup>b</sup>, Bing Zhu<sup>a</sup>, Yangai Liu<sup>a,\*</sup>

<sup>a</sup> Beijing Key Laboratory of Materials Utilization of Nonmetallic Minerals and Solid Wastes, National Laboratory of Mineral Materials, School of Material Technology, China University of Geosciences, Beijing 100083, China

<sup>b</sup> College of Materials & Environmental Engineering, Hangzhou Dianzi University, Hangzhou 310036, China

#### ARTICLE INFO

##### Keywords:

Lithium metal battery  
Lithium dendrites  
CuMOF-ANFs separator

#### ABSTRACT

Lithium metal, due to its advantages of high theoretical capacity, low density potential, is used as a negative electrode material for batteries and brings great energy storage systems. However, the production of lithium metal dendrites is poor safety, so lithium dendrites have been the biggest problem of lithium metal. The larger specific surface area and more pore structure of Cu-based metal-organic (CuMOF-ANFs) composite separator can help to inhibit the formation of lithium dendrites. The discharge capacity retention rate of the Li-Cu battery using the CuMOF-ANFs composite separator is 95% after 2000 cycles. The Li-Li batteries can continue to maintain low hysteresis for 2000 h at the 0.1 mA/cm<sup>2</sup>, the discharge capacity retention rate of the Li-Cu battery using the CuMOF-ANFs composite separator is 95% after 2000 cycles. The three-dimensional (3D) porous separator provides a new perspective for the practical application of lithium metal.

#### 1. Introduction

Certainly, here is a possible introduction for your topic: Lithium-metal batteries are promising candidates for high-energy-density rechargeable batteries due to their low electrode potentials and high theoretical capacities [1,2]. However, during the cycle, dendrites forming on the lithium metal anode can cause a short circuit, which can

chemical stability of the separator is equal to the separator remains intact and does not affect the performance of the electrolyte or other battery components. The CuMOF-ANFs composite separator helps to prevent the formation of lithium dendrites and further promote dendrite growth. Research on different materials and designs for separators with mechanical strength and chemical stability is necessary.



You

what are the issues with asking you to write papers?



ChatGPT

As an AI language model, I can certainly generate text on a wide range of topics, but there are several important considerations when it comes to asking me to write papers:

- 1. Originality:** While I can generate text, ensuring originality may be a concern. Plagiarism is unethical and can have serious consequences in academic or professional settings. Users should always verify the originality of the generated content and cite sources appropriately.
- 2. Context and Depth:** Writing papers often requires a deep understanding of the topic, context, and purpose. While I can provide information and generate text based on input, I may not fully understand the nuances of every topic or context. It's essential for users to critically assess and edit the generated content to ensure it meets the requirements of their specific task or assignment.
- 3. Quality Control:** Not all content generated by AI models may meet the desired quality standards. Users may need to review and edit the generated text to ensure coherence, accuracy, and relevance to the topic.
- 4. Ethical Considerations:** There are ethical implications to consider when using AI to generate content. Users should be transparent about the use of AI-generated text, especially in academic or professional settings where honesty and integrity are paramount.
- 5. Legal and Copyright Issues:** AI-generated content may inadvertently infringe on copyrights or intellectual property rights. Users should be aware of the legal implications and ensure that they have the right to use and distribute the generated content.

In summary, while AI language models like me can assist with writing tasks, users should approach the process with caution, ensuring originality, context relevance, and adherence to ethical and legal standards.

# Takeaways

- It is important to
  - Understand what AI can and cannot do
  - Define comprehensive set of metrics to evaluate models' performance in many aspects
- Achieving fairness in AI is a complex goal that involves addressing several key considerations:
  - **Bias Mitigation:** Striving to eliminate biases present in the training data or algorithms to prevent discriminatory outcomes. This involves identifying and rectifying any unfair treatment or disparities in how the AI system responds to different groups.
  - **Transparency:** Creating transparency in AI systems so that their decision-making processes are understandable and interpretable. Users should have insight into how decisions are made, allowing for accountability and trust.
  - **Representation:** Ensuring diverse representation in both the development teams creating AI systems and the data used to train these systems. A lack of diversity can lead to the perpetuation of biases and unfair outcomes.
  - **Equitable Access:** Making AI systems accessible and beneficial to all individuals, irrespective of their background or characteristics. This includes addressing issues related to digital divides and trying to prevent AI from exacerbating existing social inequalities.
  - **Continuous Monitoring and Improvement:** Regularly assessing and improving AI systems to identify and rectify any emerging fairness issues. This ongoing process is crucial to adapt to evolving challenges and changing societal norms.

# References

- <https://learn.microsoft.com/en-us/azure/machine-learning/concept-fairness-ml?view=azureml-api-2>
- <https://intellipaat.com/blog/what-is-uncertainty-in-artificial-intelligence/>
- [https://www.researchgate.net/figure/Illustration-of-epistemic-and-aleatoric-uncertainty\\_fig3\\_358723173](https://www.researchgate.net/figure/Illustration-of-epistemic-and-aleatoric-uncertainty_fig3_358723173)
- <https://www.weforum.org/agenda/2023/12/ai-regulation-open-source/>
- <https://www.kdnuggets.com/2023/08/shap-values-model-interpretability-machine-learning.html>
- <https://www.relativity.com/blog/the-white-house-addresses-responsible-ai-eo-takeaways-on-fairness/#:~:text=Achieving%20fairness%20in%20AI%20is,algorithms%20to%20prevent%20discriminatory%20outcomes.>