Artificial Intelligence
Advanced Topics in AI & ML

Interpretability, Explainability, and AI Ethics

Aleksandr Petiushko

ML Research

SOFIA UNIVERSITY

AP

Content

1. Interpretability

Content

1. Interpretability
2. Explainability

Content

1. Interpretability
2. Explainability
3. Bias and Fairness in AI

Content

1. Interpretability
2. Explainability
3. Bias and Fairness in AI
4. AI Ethics

Content

1. Interpretability
2. Explainability
3. Bias and Fairness in AI
4. AI Ethics
5. AI Regulations

Interpretability

- Interpretability: understand the influence of any input sub-area/sub-feature on the model output;

Interpretability

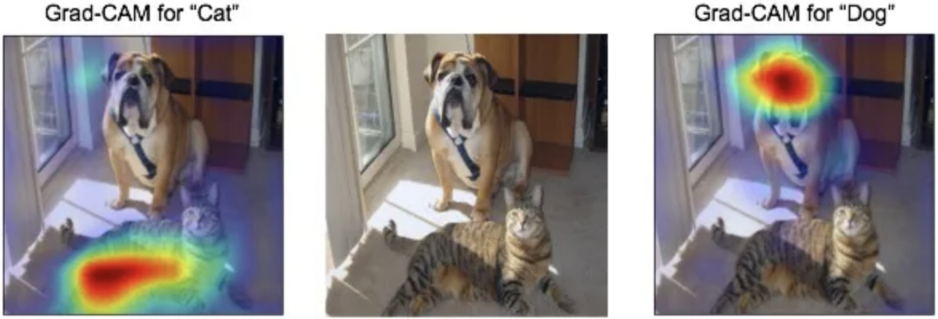
- Interpretability: understand the influence of any input sub-area/sub-feature on the model output;
- Can be understood as a sophisticated tool towards the ML Debug system

Interpretability

- Interpretability: understand the influence of any input sub-area/sub-feature on the model output;
- Can be understood as a sophisticated tool towards the ML Debug system
- Can be done via input counterfactual analysis (changing/reverting some input features)

Interpretability

- Interpretability: understand the influence of any input sub-area/sub-feature on the model output;
- Can be understood as a sophisticated tool towards the ML Debug system
- Can be done via input counterfactual analysis (changing/reverting some input features)
- Read material: link



Explainability

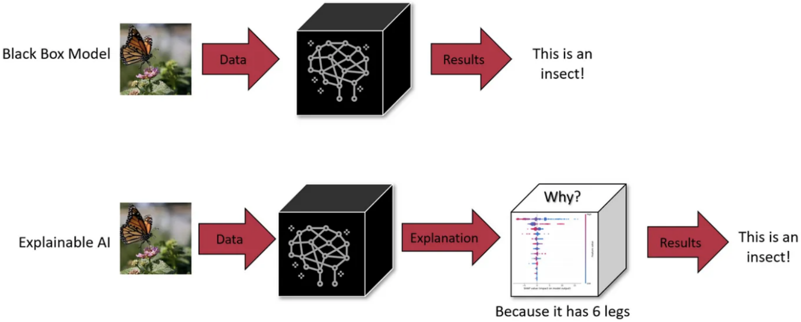
- Explainability: high-level interpretability (using human-like language), a clear and intuitive explanation of the decisions made

Explainability

- Explainability: high-level interpretability (using human-like language), a clear and intuitive explanation of the decisions made
- Explainability now can be done via LLM chain-of-thought (CoT) technique

Explainability

- Explainability: high-level interpretability (using human-like language), a clear and intuitive explanation of the decisions made
- Explainability now can be done via LLM chain-of-thought (CoT) technique
- Read material: link old, link new



Bias and Fairness in AI

- Fairness is the subjective practice of using AI without favoritism or discrimination, and Bias is the preference or prejudice against a feature (so roughly speaking when talking about people they are almost synonyms)

Bias and Fairness in AI

- Fairness is the subjective practice of using AI without favoritism or discrimination, and Bias is the preference or prejudice against a feature (so roughly speaking when talking about people they are almost synonyms)
- ML model can provide biased predictions w.r.t. race, gender, age. etc

Bias and Fairness in AI

- Fairness is the subjective practice of using AI without favoritism or discrimination, and Bias is the preference or prejudice against a feature (so roughly speaking when talking about people they are almost synonyms)
- ML model can provide biased predictions w.r.t. race, gender, age. etc
- The main reason: skewed/sparse training data

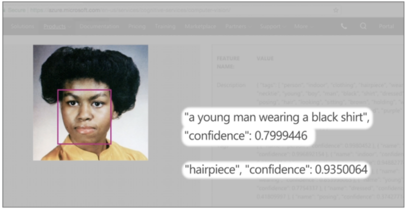
Bias and Fairness in AI

- Fairness is the subjective practice of using AI without favoritism or discrimination, and Bias is the preference or prejudice against a feature (so roughly speaking when talking about people they are almost synonyms)
- ML model can provide biased predictions w.r.t. race, gender, age. etc
- The main reason: skewed/sparse training data
- Main technique to avoid: human-in-the-loop, alignment

Bias and Fairness in AI

- Fairness is the subjective practice of using AI without favoritism or discrimination, and Bias is the preference or prejudice against a feature (so roughly speaking when talking about people they are almost synonyms)
- ML model can provide biased predictions w.r.t. race, gender, age. etc
- The main reason: skewed/sparse training data
- Main technique to avoid: human-in-the-loop, alignment
- Read material: link old, link new

Michelle Obama

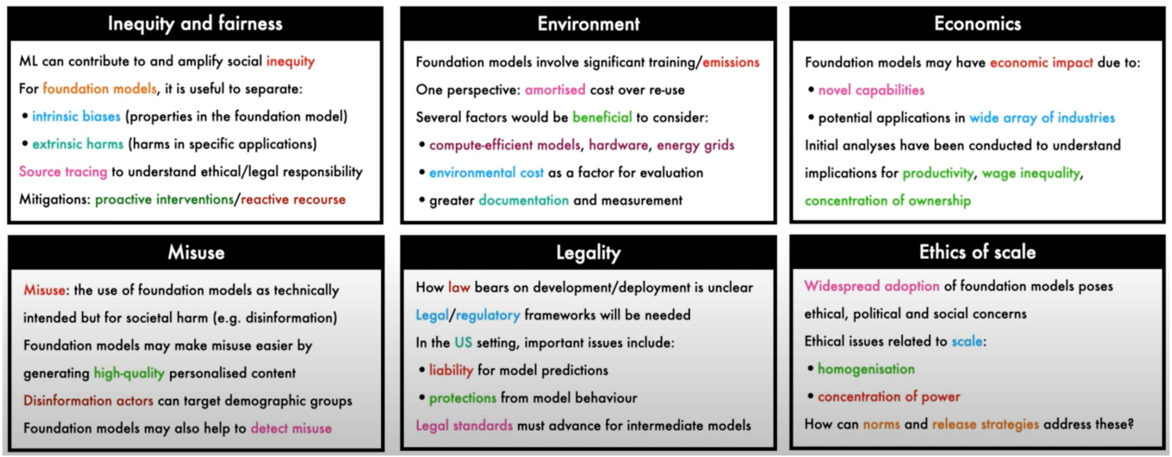


"a young man wearing a black shirt",
"confidence": 0.7999446

"hairpiece", "confidence": 0.9350064

Microsoft

AI Ethics[1]



Inequity and fairness

ML can contribute to and amplify social inequity

For foundation models, it is useful to separate:

- intrinsic biases (properties in the foundation model)
- extrinsic harms (harms in specific applications)

Source tracing to understand ethical/legal responsibility

Mitigations: proactive interventions/reactive recourse

Environment

Foundation models involve significant training/emissions

One perspective: amortised cost over re-use

Several factors would be beneficial to consider:

- compute-efficient models, hardware, energy grids
- environmental cost as a factor for evaluation
- greater documentation and measurement

Economics

Foundation models may have economic impact due to:

- novel capabilities
- potential applications in wide array of industries

Initial analyses have been conducted to understand implications for productivity, wage inequality, concentration of ownership

Misuse

Misuse: the use of foundation models as technically intended but for societal harm (e.g. disinformation)

Foundation models may make misuse easier by generating high-quality personalised content

Disinformation actors can target demographic groups

Foundation models may also help to detect misuse

Legality

How law bears on development/deployment is unclear

Legal/regulatory frameworks will be needed

In the US setting, important issues include:

- liability for model predictions
- protections from model behaviour

Legal standards must advance for intermediate models

Ethics of scale

Widespread adoption of foundation models poses ethical, political and social concerns

Ethical issues related to scale:

- homogenisation
- concentration of power

How can norms and release strategies address these?

[1] www.youtube.com

AI Regulations 2024

- Main directions of regulations:
    1. Ensuring non-biased and privacy-concerned AI solutions,

AI Regulations 2024

- Main directions of regulations:
  1. Ensuring non-biased and privacy-concerned AI solutions,
  2. Fears of “Harmful” (potentially dangerous) AI;

AI Regulations 2024

- Main directions of regulations:
  1. Ensuring non-biased and privacy-concerned AI solutions,
  2. Fears of “Harmful” (potentially dangerous) AI;
- CA SB (California Senate Bill) 1047 (Feb 2024)

AI Regulations 2024

- Main directions of regulations:
  1. Ensuring non-biased and privacy-concerned AI solutions,
  2. Fears of “Harmful” (potentially dangerous) AI;
- CA SB (California Senate Bill) 1047 (Feb 2024)
  - Bans (actually, requires a “kill switch”) AI if it uses more than 10^{26} FLOPs / $100M for training or more than 10^{25} FLOPs / $10M for finetuning

AI Regulations 2024

- Main directions of regulations:
  1. Ensuring non-biased and privacy-concerned AI solutions,
  2. Fears of “Harmful” (potentially dangerous) AI;
- CA SB (California Senate Bill) 1047 (Feb 2024)
  - Bans (actually, requires a “kill switch”) AI if it uses more than 10^{26} FLOPs / \$100M for training or more than 10^{25} FLOPs / \$10M for finetuning
  - Governor of CA (Gavin Newsom) vetoes it (Sep 2024)

AI Regulations 2024

- Main directions of regulations:
  1. Ensuring non-biased and privacy-concerned AI solutions,
  2. Fears of "Harmful" (potentially dangerous) AI;
- CA SB (California Senate Bill) 1047 (Feb 2024)
  - Bans (actually, requires a "kill switch") AI if it uses more than 10^{26} FLOPs / \$100M for training or more than 10^{25} FLOPs / \$10M for finetuning
  - Governor of CA (Gavin Newsom) vetoes it (Sep 2024)
- European Union's AI Act (Aug 2024) restricts AI in essential fields like education, employment, and law enforcement, and requires developers to provide details of their algorithms and data

AI Regulations 2024

- Main directions of regulations:
  1. Ensuring non-biased and privacy-concerned AI solutions,
  2. Fears of “Harmful” (potentially dangerous) AI;
- CA SB (California Senate Bill) 1047 (Feb 2024)
  - Bans (actually, requires a “kill switch”) AI if it uses more than 10^{26} FLOPs / \$100M for training or more than 10^{25} FLOPs / \$10M for finetuning
  - Governor of CA (Gavin Newsom) vetoes it (Sep 2024)
- European Union’s AI Act (Aug 2024) restricts AI in essential fields like education, employment, and law enforcement, and requires developers to provide details of their algorithms and data
- Framework Convention on AI and Human Rights, Democracy, and the Rule of Law was signed by the US, UK, European and other countries (Sep 2024)

AI Regulations 2024

- Main directions of regulations:
  1. Ensuring non-biased and privacy-concerned AI solutions,
  2. Fears of “Harmful” (potentially dangerous) AI;
- CA SB (California Senate Bill) 1047 (Feb 2024)
  - Bans (actually, requires a “kill switch”) AI if it uses more than 10^{26} FLOPs / \$100M for training or more than 10^{25} FLOPs / \$10M for finetuning
  - Governor of CA (Gavin Newsom) vetoes it (Sep 2024)
- European Union’s AI Act (Aug 2024) restricts AI in essential fields like education, employment, and law enforcement, and requires developers to provide details of their algorithms and data
- Framework Convention on AI and Human Rights, Democracy, and the Rule of Law was signed by the US, UK, European and other countries (Sep 2024)
  - Requires AI models respect democracy and human rights

AI Regulations 2025

- Governor of CA (Gavin Newsom) signed:
  - CA SB 53 (September 2025): requirement for AI companies if their model uses more than 10^{26} FLOPs / annual revenue above \$500M to provide transparency about capabilities and potential risks, and publish safety frameworks

AI Regulations 2025

- Governor of CA (Gavin Newsom) signed:
  - CA SB 53 (September 2025): requirement for AI companies if their model uses more than 10^{26} FLOPs / annual revenue above $500M to provide transparency about capabilities and potential risks, and publish safety frameworks
  - CA (Assembly Bill) SB 243 (October 2025): requirement to prevent chatbots from harming minors and other vulnerable users

AI Regulations 2025

- Governor of CA (Gavin Newsom) signed:
  - CA SB 53 (September 2025): requirement for AI companies if their model uses more than 10^{26} FLOPs / annual revenue above $500M to provide transparency about capabilities and potential risks, and publish safety frameworks
  - CA (Assembly Bill) SB 243 (October 2025): requirement to prevent chatbots from harming minors and other vulnerable users
  - CA AB 316 (October 2025): makes developers liable for the actions of AI autonomous systems they build

AI Regulations 2025

- Governor of CA (Gavin Newsom) signed:
  - CA SB 53 (September 2025): requirement for AI companies if their model uses more than 10^{26} FLOPs / annual revenue above $500M to provide transparency about capabilities and potential risks, and publish safety frameworks
  - CA (Assembly Bill) SB 243 (October 2025): requirement to prevent chatbots from harming minors and other vulnerable users
  - CA AB 316 (October 2025): makes developers liable for the actions of AI autonomous systems they build
  - CA AB 853 (October 2025): requirement AI-generated media to be labeled clearly

AI Regulations 2025

- Governor of CA (Gavin Newsom) signed:
  - CA SB 53 (September 2025): requirement for AI companies if their model uses more than 10^{26} FLOPs / annual revenue above $500M to provide transparency about capabilities and potential risks, and publish safety frameworks
  - CA (Assembly Bill) SB 243 (October 2025): requirement to prevent chatbots from harming minors and other vulnerable users
  - CA AB 316 (October 2025): makes developers liable for the actions of AI autonomous systems they build
  - CA AB 853 (October 2025): requirement AI-generated media to be labeled clearly
- Governor of CA (Gavin Newsom) vetoed:
  - CA SB 7 (October 2025): requirement to notify if employers used AI for hiring/firing

AI Regulations 2025

- Governor of CA (Gavin Newsom) signed:
  - CA SB 53 (September 2025): requirement for AI companies if their model uses more than 10^{26} FLOPs / annual revenue above \$500M to provide transparency about capabilities and potential risks, and publish safety frameworks
  - CA (Assembly Bill) SB 243 (October 2025): requirement to prevent chatbots from harming minors and other vulnerable users
  - CA AB 316 (October 2025): makes developers liable for the actions of AI autonomous systems they build
  - CA AB 853 (October 2025): requirement AI-generated media to be labeled clearly
- Governor of CA (Gavin Newsom) vetoed:
  - CA SB 7 (October 2025): requirement to notify if employers used AI for hiring/firing

Takeaway notes

1. Read all the mentioned links

Takeaway notes

1. Read all the mentioned links
2. Interpretability and Explainability are quite connected in ML

Takeaway notes

1. Read all the mentioned links
2. Interpretability and Explainability are quite connected in ML
3. Interpretability deals mostly on a lower level, input/output dependencies

Takeaway notes

1. Read all the mentioned links
2. Interpretability and Explainability are quite connected in ML
3. Interpretability deals mostly on a lower level, input/output dependencies
4. Explainability steps in on a higher level to provide a human-like explanations

Takeaway notes

1. Read all the mentioned links
2. Interpretability and Explainability are quite connected in ML
3. Interpretability deals mostly on a lower level, input/output dependencies
4. Explainability steps in on a higher level to provide a human-like explanations
5. Usually the most interpretable are simpler models; explainability can be applied to a model of any complexity

Thank you!