Brookhaven National Laboratory

U.S. DEPARTMENT OF ENERGY

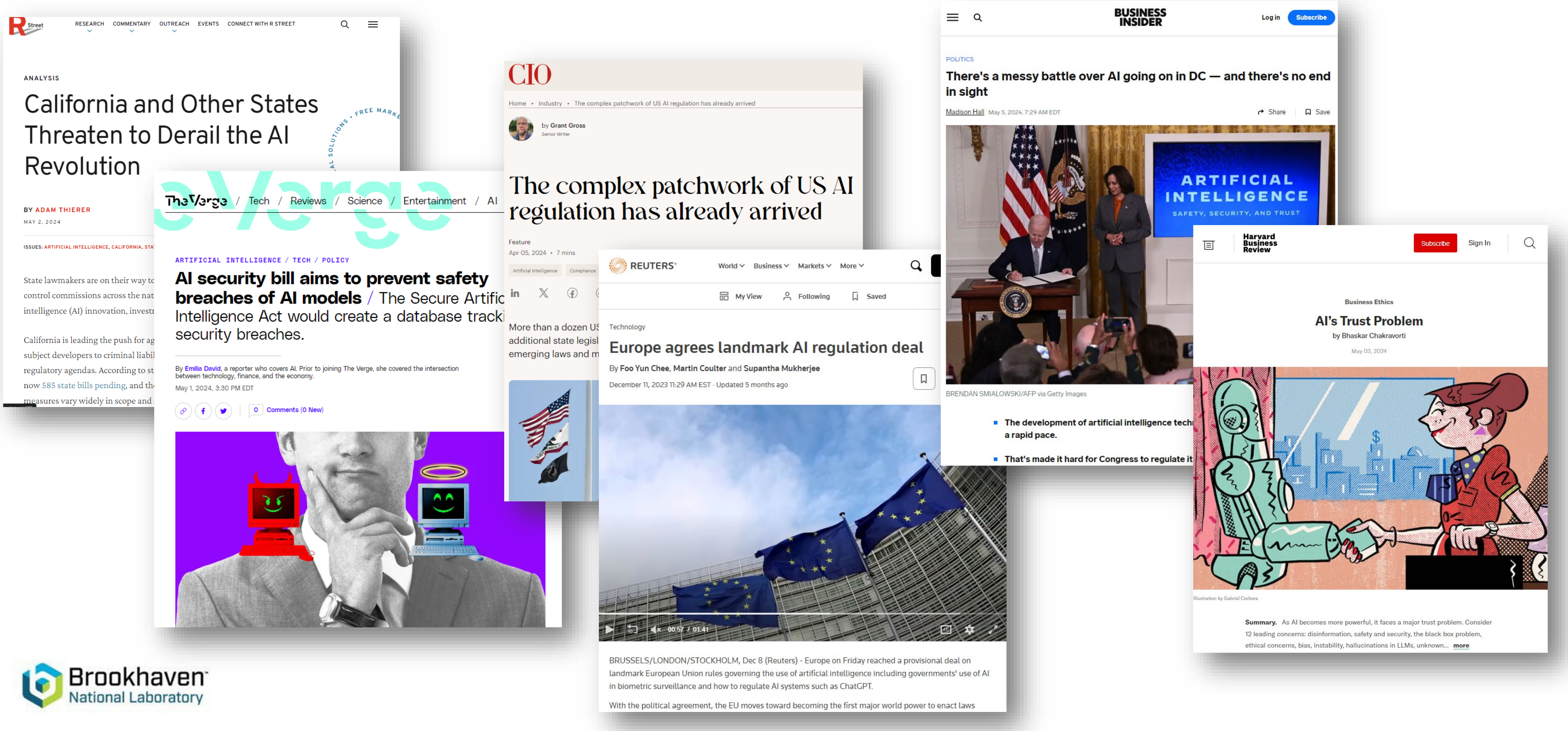
# Safe, Secure and Trustworthy AI

Carlos Soto

AI Theory and Security Group

March 7, 2024

# Safe, Secure and Trustworthy AI...



ANALYSIS

California and Other States Threaten to Derail the AI Revolution

BY ADAM THIERER
MAY 2, 2024

ISSUES: ARTIFICIAL INTELLIGENCE, CALIFORNIA, STA

State lawmakers are on their way to
control commissions across the nat
intelligence (AI) innovation, investr

California is leading the push for ag
subject developers to criminal liabil
regulatory agendas. According to st
now 585 state bills pending, and th
measures vary widely in scope and

ARTIFICIAL INTELLIGENCE / TECH / POLICY

**AI security bill aims to prevent safety breaches of AI models** / The Secure Artific
Intelligence Act would create a database tracki
security breaches.

By Emilia David, a reporter who covers AI. Prior to joining The Verge, she covered the intersection between technology, finance, and the economy.
May 1, 2024, 3:30 PM EDT

CIO
Home • Industry • The complex patchwork of US AI regulation has already arrived

by Grant Gross
Senior Writer

The complex patchwork of US AI regulation has already arrived

Feature
Apr 05, 2024 • 7 mins

More than a dozen US
additional state legisl
emerging laws and m

Technology

Europe agrees landmark AI regulation deal

By Foo Yun Chee, Martin Coulter and Supantha Mukherjee

December 11, 2023 11:29 AM EST · Updated 5 months ago

BRUSSELS/LONDON/STOCKHOLM, Dec 8 (Reuters) - Europe on Friday reached a provisional deal on landmark European Union rules governing the use of artificial intelligence including governments' use of AI in biometric surveillance and how to regulate AI systems such as ChatGPT.

With the political agreement, the EU moves toward becoming the first major world power to enact laws

POLITICS

There's a messy battle over AI going on in DC — and there's no end in sight

Madison Hall May 5, 2024, 7:29 AM EDT

BRENDAN SMIALOWSKI/AFP via Getty Images

- The development of artificial intelligence tech
  a rapid pace.
- That's made it hard for Congress to regulate it

Business Ethics

AI's Trust Problem

by Bhaskar Chakravorti

May 03, 2024

Illustration by Gabriel Corbera

**Summary.** As AI becomes more powerful, it faces a major trust problem. Consider 12 leading concerns: disinformation, safety and security, the black box problem, ethical concerns, bias, instability, hallucinations in LLMs, unknown... **more**

# Safe, Secure and Trustworthy AI

- AI here and now
- Robustness
- Privacy and Security
- What about generative AI?
- Transparency and ecosystems

# It’s less about Skynet and HAL-9000..

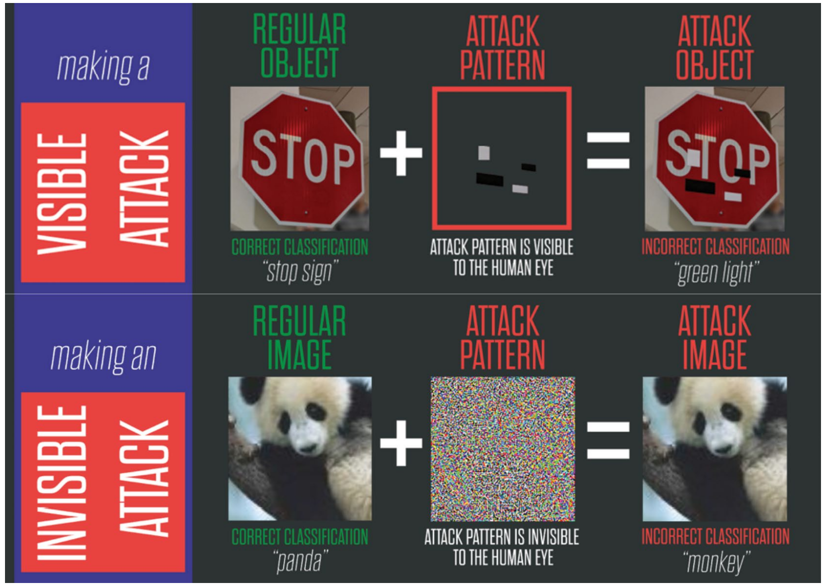
.. than about how AI is used today.

- How AI models are built
- How they are used
- How people *want* to use them
- What to expect from AI
- What you may not consider

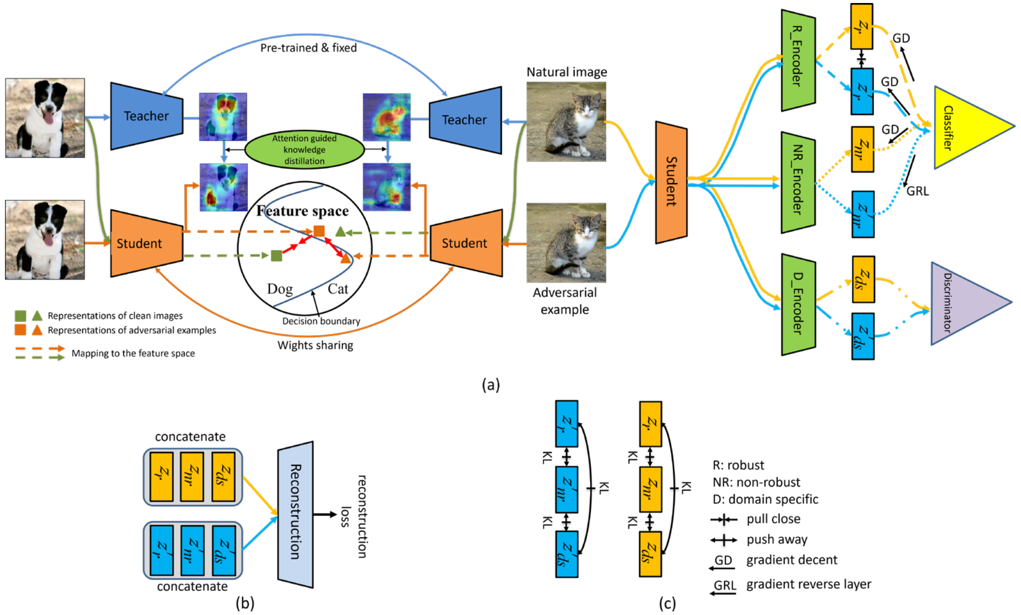
# Robustness and Consistency

AI models that “work” may still fail in the real world

- Error sources: noise, class overlap, shift in data distribution, **adversarial attacks**
  - Some attacks require knowledge of model details (*white-box*), other don’t (*black-box*)
- Robust training of AI models can *reduce* susceptibility to attacks


making a
VISIBLE ATTACK
REGULAR OBJECT
ATTACK PATTERN
ATTACK OBJECT
CORRECT CLASSIFICATION "stop sign"
ATTACK PATTERN IS VISIBLE TO THE HUMAN EYE
INCORRECT CLASSIFICATION "green light"
making an
INVISIBLE ATTACK
REGULAR IMAGE
ATTACK PATTERN
ATTACK IMAGE
CORRECT CLASSIFICATION "panda"
ATTACK PATTERN IS INVISIBLE TO THE HUMAN EYE
INCORRECT CLASSIFICATION "monkey"

**Adversarial attacks on AI model** (*Kotyan 2023)*



**Robust AI training to counter adversarial attacks** *(Hong Wang, BNL 2024)*

# Confidence and Uncertainty

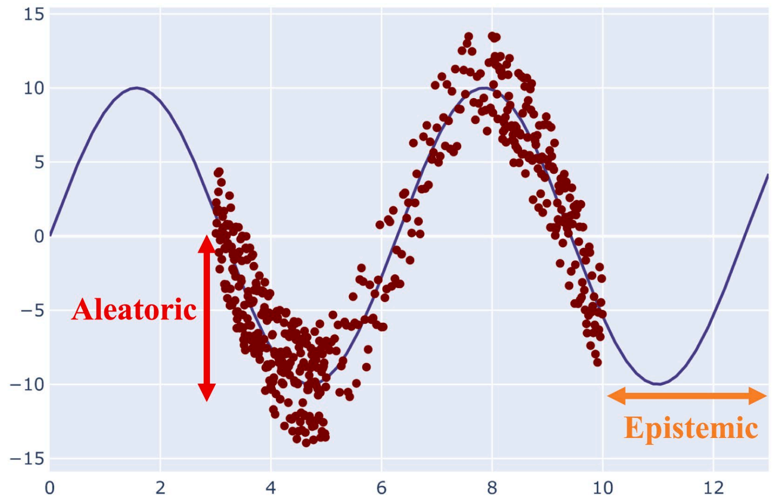
Closely related to robustness (confidence in AI responses, consistency)

Two types of uncertainty:

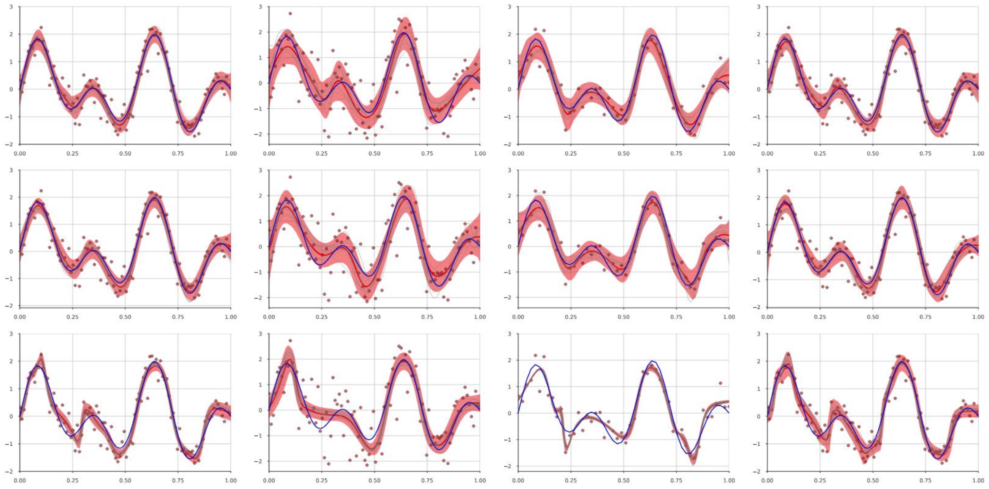
- **Aleatoric:** data uncertainty (e.g. from measurements), induces bias, propagates through model, *irreducible* without additional data
- **Epistemic**: knowledge/modeling uncertainty, *may be reduced* with improved modeling

Techniques such as Bayesian inference may directly estimate AI uncertainty & robustness

- Challenging to extend to large models



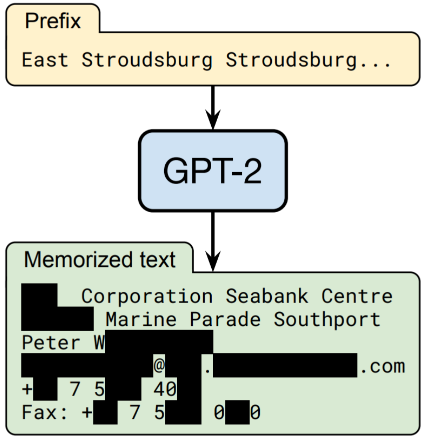
**Data (aleatoric) and modeling (epistemic) uncertainty**
(*Abdar 2021*)



**Bayesian inference for uncertainty quantification**
(*Sanket Jantre, BNL 2023*)

# Private and Secure AI Training and Inference

- Training on private/sensitive data poses challenge
  - Training data leakage
  - Moving sensitive data to AI compute
- Privacy-preserving AI
  - **Differential privacy (DP)** introduces stochastic noise to mask individual data samples
  - BNL demonstrated first distributed DP
- Secure AI training and inference
  - **Fully Homomorphic Encryption (FHE)** enables training without ever exposing secure data



Prefix

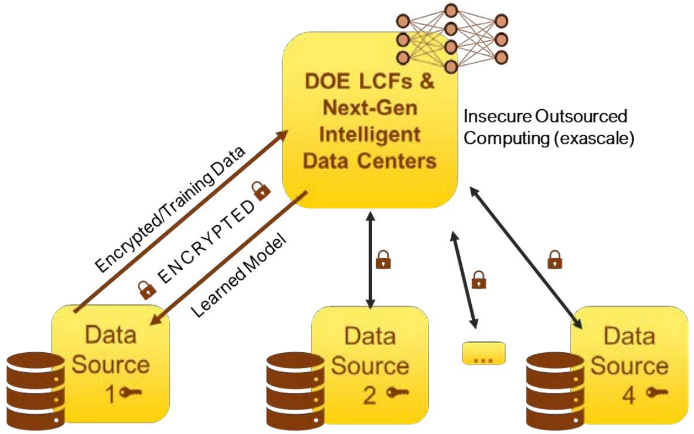
East Stroudsburg Stroudsburg...

GPT-2

Memorized text

[redacted] Corporation Seabank Centre
Marine Parade Southport
Peter W[redacted]
[redacted]@[redacted].[redacted].com
+[redacted] 7 5[redacted] 40[redacted]
Fax: +[redacted] 7 5[redacted] 0[redacted]0

**Extracting training data from deployed model**
*(Carlini 2021)*

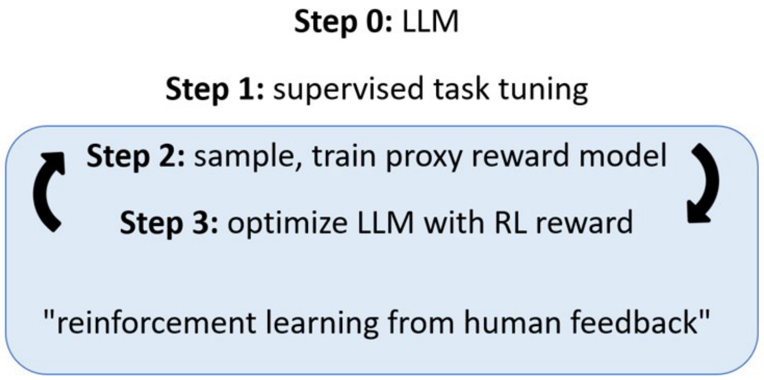


**Using FHE to train secure AI models on encrypted data**

# Safety, Security and Trust for Generative AI

Large Language Models (LLMs) and other generative Foundation Models (FMs) pose additional challenges

- Biases and alignment issues
  - May be mitigated *or reinforced* by **RLHF (Reinforcement Learning from Human Feedback)**
  - e.g., personification
- Hallucinations and Verification
  - Imperfect memorization, pseudo-reasoning
  - May be mitigated in part by **Chain-of-Thought (CoT)** reasoning, **self-critique**, **multi-LLM ensembles**, and leveraging external resources, e.g., **Retrieval-Augmented Generation (RAG)**
- Other risks: data poisoning, prompt injections



**Step 0:** LLM

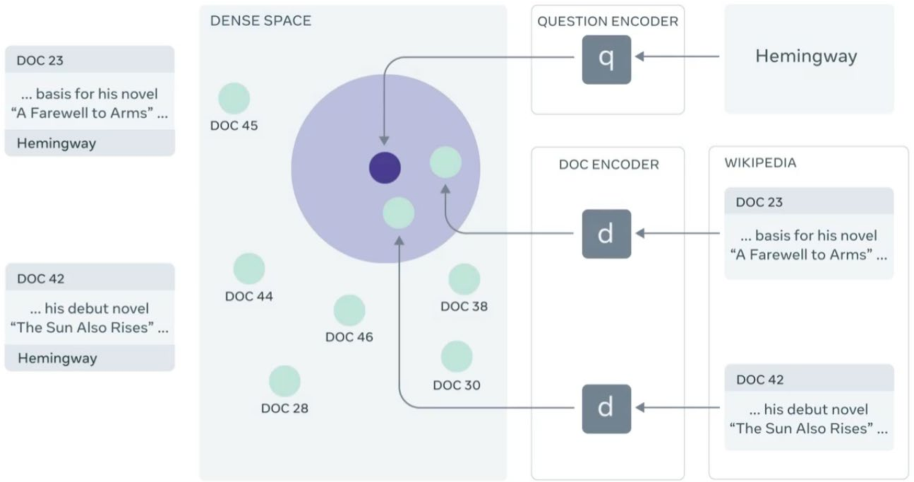
**Step 1:** supervised task tuning

**Step 2:** sample, train proxy reward model

**Step 3:** optimize LLM with RL reward

"reinforcement learning from human feedback"

**RLHF is used to align LLM output with human reviewer expectations**


DENSE SPACE
QUESTION ENCODER
q
Hemingway
DOC 23
... basis for his novel "A Farewell to Arms" ...
Hemingway
DOC 45
DOC ENCODER
WIKIPEDIA
DOC 23
... basis for his novel "A Farewell to Arms" ...
d
DOC 42
... his debut novel "The Sun Also Rises" ...
Hemingway
DOC 44
DOC 38
DOC 46
DOC 30
DOC 28
d
DOC 42
... his debut novel "The Sun Also Rises" ...

**RAG can map queries and knowledge resources to common embedding space, LLM retrieves relevant context to inject in prompt or response** *(Lewis 2020)*

Brookhaven National Laboratory

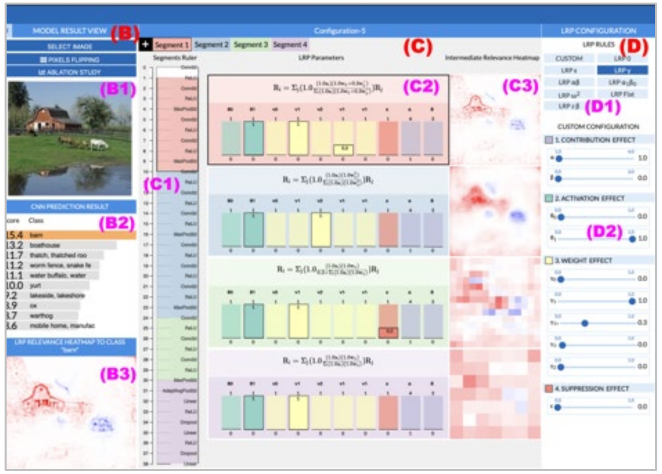
# Transparency and AI Ecosystems

Interpretable and **Explainable AI (XAI)** continues to grow in relevance

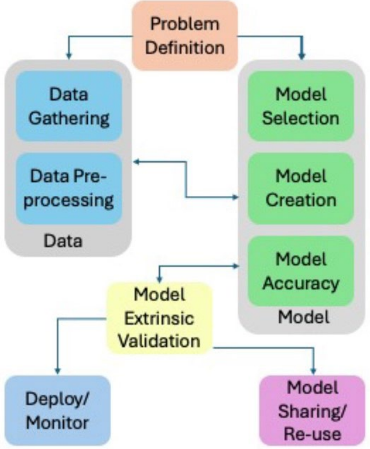
- Feature and activation **visualizations** enable per-instance inspection
- Model-agnostic **interpretation** techniques (e.g. LIME, ICE, ALE, SHAP) generally aim to identify feature contributions

Trust, safety, and security of AI ultimately comes down to **use**

- Risks exist throughout **AI ecosystems and lifecycle**



**Layer-wise Relevance Propagation (LRP) visualization for deep NNs**
*(Huang 2021)*


Problem Definition

Data Gathering

Data Pre-processing

Data

Model Selection

Model Creation

Model Accuracy

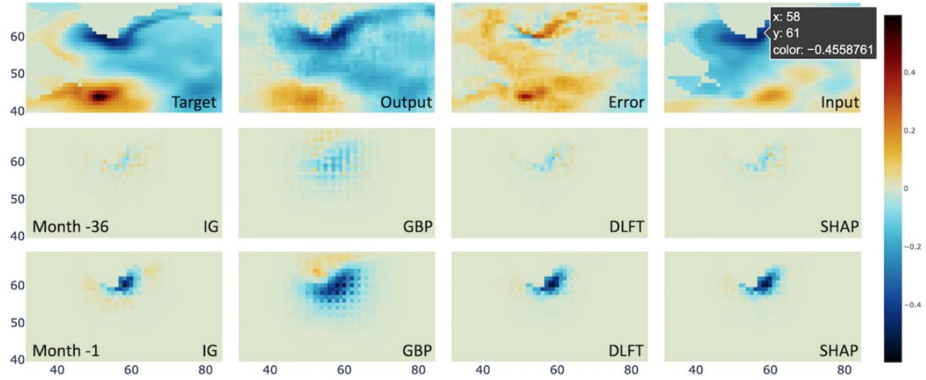
Model

Model Extrinsic Validation

Deploy/ Monitor

Model Sharing/ Re-use

**The AI Lifecycle**
*(Castro 2023)*



**Visualizing feature importance in AI climate model**
*(Wei Xu, BNL 2021)*

Brookhaven National Laboratory

# Thank you

csoto@bnl.gov

## References

- Kotyan, Shashank. "A reading survey on adversarial machine learning: Adversarial attacks and their understanding." *arXiv preprint arXiv:2308.03363* (2023).
- Wang, Hong, et al. "Exploring robust features for improving adversarial robustness." *IEEE Transactions on Cybernetics* (2024).
- Abdar, Moloud, et al. "A review of uncertainty quantification in deep learning: Techniques, applications and challenges." *Information fusion* 76 (2021): 243-297.
- Jantre, Sanket, et al. "Learning active subspaces for effective and scalable uncertainty quantification in deep neural networks." *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024.
- Carlini, Nicholas, et al. "Extracting training data from large language models." *30th USENIX Security Symposium (USENIX Security 21)*. 2021.
- Huang, Xinyi, et al. "A Visual Designer of Layer-wise Relevance Propagation Models." *Computer Graphics Forum*. Vol. 40. No. 3. 2021.
- Xu, Wei, et al. "Feature importance in a deep learning climate emulator." *Modeling Oceans and Climate Change Workshop at ICLR* 2021.
- Castro, Leyla Jael, et al. “Lifecycle for FAIR Machine Learning.” *technical report*, 2023

Brookhaven National Laboratory