Safe, Secure and Trustworthy AI

Carlos Soto
AI Theory and Security Group

March 7, 2024
Safe, Secure and Trustworthy AI...
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
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
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
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

---

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

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

The AI Lifecycle (Castro 2023)

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

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

csoto@bnl.gov

References