Advancing Secure, Trustworthy, and Energy-Efficient AI for Science and Technology: A view from ORNL’s AI Initiative

Prasanna Balaprakash
Director of AI Programs
Oak Ridge National Laboratory
Solar Applications of Artificial Intelligence and Machine Learning Workshop
October 31, 2023
DOE and ORNL mission AI applications
Accelerating scientific discovery, fortifying energy infrastructure, and enhancing national security
Grand challenges in AI for science and security
Paradox of AI development and challenges

- Easy to demo but hard in production
- Hard problems are easy and the easy problems are hard
- Ever growing open research problems
- Humans remain a roadblock
- Unique challenges with cyber-physical systems
Paradox of efficiency: Goodhart’s law

Too much efficiency makes everything worse

Well-aligned phase

Overfitting / Goodhart’s law

https://sohl-dickstein.github.io/2022/11/06/strong-Goodhart.html
Alignment

**Alignment**: Ensuring that AI systems’ goals and behaviors align with science and human values and intentions

**Importance**: Prevent potential harmful consequences of AI actions that could result from misalignment

**Challenges**: Defining human values, transferring these values to AI, and allowing for value learning and adaptation over time

**Continuous Effort**: Continuous effort as AI evolves and as societal values change
Driving safely on the road to AI implementation: Guardrails for responsible AI use

**Destination (Objective):** Effective Decision Making, Predictive Analysis, Automated Operations, and Improved Efficiency

**Obstacles (Challenges):** Bias, Misuse, Lack of Understanding, Complexity

**Guardrails (Safety measures):** Ethics, Transparency, Privacy, Fairness, Security
Quality assurance in AI: Ensuring we're not only building the AI product right but also building the right AI product
Moving beyond correlations: Causal modeling for predictive and explanatory power

uncover why something happens, not just what happens next
ORNL’s AI initiative
Secure, trustworthy, and energy-efficient AI

AI for scientific discovery and complex systems
- Secure
  - Alignment
  - Cybersecurity
  - Robustness

AI for experimental facilities
- Trustworthy
  - Validation and verification
  - Uncertainty qualifications
  - Casual reasoning

AI for national security
- Energy efficient
  - Scalability
  - Edge
  - Co-design
Safe AI: Goal and behavior alignment with science, human values, and intentions

Considerations

- Accuracy
- Fairness
- Privacy
- Transparency
- Robustness
- Energy-efficiency
CAISER – Center for AI SEcurity Research

National center of excellence with strong leadership:
Leading-edge NS programs
AI initiative
Computing excellence
Computing resources

Safeguarding AI systems against threats

- Safeguarding AI data and models from unauthorized access
- Consistent monitoring and auditing of AI operations and frameworks
- Understanding and addressing data and model poisoning
- Establishment of mitigation strategies (Secure data management and robust training methodologies)
Assurance: Reliable, Robust, and Safe AI

Uncertainty Quantification (UQ)

Verification & Validation (V&V)

Explainability & Interpretability

Privacy

Uncertainty Quantification for Trustworthy AI

- Neutron sciences
  - Neutron diffraction data analysis
- Natural science
  - Prediction of streamflow, temperature, carbon flux
- Smart grid system
  - Prediction of transient source locations
- R-CNN
- FNN
- PNN
- RNN
- CNN
- GCN
- GWN

Privacy-Preserving Model Training

- Train ML on Private Data
  - FNN: Forward NN
  - RNN: Recurrent NN
  - CNN: Convolutional NN
  - GWN: GraphWaveNet
  - GCN: Graph convolutional network
  - R-CNN: Regional-CNN

Privacy-Preservation at Edge

- Automatic privacy-preservation of streaming data on edge such as smart grid
- train and release ML models on a private dataset with a formal privacy guarantee

reliable and scalable uncertainty quantification methods for DOE mission area
Validation and verification

• Sampling-based approach to quantitatively estimate properties for deep neural networks (DNN) with probabilistic guarantees
  – Given a logical property $\psi$ specified over a space of inputs and outputs of a DNN and a numerical threshold $\theta$, decide whether $\psi$ is true for less than $\theta$ fraction of the inputs
  – Assumes only black box access
  – Provides quantitative verification of properties like fairness, privacy, and robustness
  – Verification is sound – when $\psi$ is confirmed to be true, it can be deduced mathematically

Dataset $f(x)$

\[
N = \frac{(\sqrt{3q_1 + \sqrt{2q_2}})^2}{(q_2 - q_1)^2} \ln \frac{1}{\delta}
\]

Samples

DNN

Labels

Property Verified?

Provero

Robustness Verification

Compare $p$, $\theta$

Adversarial Density, $p$

$p$ is the probability that $\psi$ evaluates to 1 for a given sample
$\theta$ is the threshold
ORNL’s AI initiative
Secure, trustworthy, and energy-efficient AI

The AI Initiative leverage and enhance ORNL’s existing facilities and capabilities

- ORNL Center for AI Security Research (CAISER)
- INTERSECT
- Secure, Trustworthy, and Energy-Efficient AI
- OLCF
- CITADEL
AI workshop series

ORNL's Generative AI Workshop Series:
2nd Workshop Towards Safe, Trustworthy, and Energy-Efficient AI Models

Co-located with the Smoky Mountain Conference 2023
Tuesday, August 29th 2023
Crowne Plaza Hotel, Knoxville Downtown, TN
Hybrid event
ORNL’s AI initiative
Secure, trustworthy, and energy-efficient AI

AI for scientific discovery and complex systems
- Secure
  - Alignment
  - Cybersecurity
  - Robustness
- Trustworthy
  - Validation and verification
  - Uncertainty qualifications
  - Casual reasoning
- Energy efficient
  - Scalability
  - Edge
  - Co-design

AI for experimental facilities
AI for national security