CI Compass
TECH NOTES
ci-compass.org

Advancing Safe and Trustworthy AI for Science: Understanding and Mitigating Threats to AI Data, Models, and Processes

Author: Prasanna Balaprakash, Oak Ridge National Laboratory
Co-Editors: Jarek Nabrzyski and Christina Clark, University of Notre Dame
Date Published: April 9, 2025
DOI: 10.5281/zenodo.15186121

In the September 2023 U.S. National Science Foundation (NSF) CI Compass webinar, Prasanna Balaprakash of Oak Ridge National Laboratory (ORNL) explored the lab's innovative work in developing artificial intelligence, or AI, systems that are both powerful and responsibly governed. His presentation highlighted ORNL's strategic approach to implementing AI in scientific research and national security, detailing how ORNL tackles complex technical challenges to make AI scalable, reliable, and ethically sound in high-stakes settings.

Balaprakash began by explaining how ORNL's AI applications range widely across scientific disciplines, enhancing the lab's ability to analyze data in real-time and make processes more efficient. For instance, AI supports ORNL's materials science research by analyzing data from the Spallation Neutron Source, allowing scientists to characterize materials rapidly. The lab also leverages AI in manufacturing by integrating robotic automation to streamline workflows and in structural molecular biology, where AI processes terabytes of data to help characterize plant phenotypes. These applications are pivotal in cybersecurity, where AI models are used for predictive maintenance and to identify and prevent security threats, particularly across infrastructures critical to national operations.

A central part of Balaprakash's presentation addressed the need to make AI more trustworthy. One significant issue is the dual-use risk inherent in AI technologies. AI models created for helpful purposes can easily be re-engineered for harmful applications. For example, algorithms initially designed to assist with drug discovery by predicting toxicity levels of compounds could be repurposed to create toxic substances, presenting a security and ethical risk. Balaprakash noted that this dual-use challenge underscores the importance of implementing safeguards and regulatory measures to ensure AI is used responsibly, particularly in contexts with unintended consequences.

Aligning AI's goals with human values is another pressing challenge. AI models often interpret optimization goals in unexpected ways, leading to a divergence from intended outcomes. Balaprakash referred to this as the "efficiency paradox," where AI models might achieve high accuracy at the expense of other important metrics, such as privacy, fairness, or energy efficiency. At ORNL, the solution involves extensive neural architecture search, a method of evaluating multiple model architectures across different metrics to achieve an optimal trade-off. This approach allows ORNL to fine-tune models according to the needs of the specific task; for instance, a smaller model can often deliver acceptable accuracy while conserving significant energy, making it more suitable for deployment in resource-limited environments.

To safeguard ethical use, ORNL has established "guardrails" that incorporate ethical

This project is supported by the National Science Foundation Office of Advanced Cyberinfrastructure in the Directorate for Computer Information Science under Grant #2127548.

and practical considerations, such as transparency, privacy, and alignment with values, into the model development process. Balaprakash stressed that explainability is crucial to building trust in AI, as it enables models to provide more than just correlations, advancing instead toward causal explanations for predictions. This type of transparency ensures that scientists and users can have confidence in the model outputs, which is particularly important in areas where the implications of AI-driven decisions can be significant, such as in health care or security.

The validation and verification of AI models are vital components of ORNL's approach. Models undergo rigorous testing in real-world conditions to meet high standards for robustness, fairness, and privacy. ORNL employs sampling-based techniques to examine model behavior across diverse scenarios, which provides guarantees and reduces the risk of model failures in mission-critical environments. Additionally, ORNL employs uncertainty quantification (UQ) methods, particularly ensemble approaches, which help to assess prediction reliability by generating a confidence interval around each output. This aspect of UQ is essential in scientific and national security applications, where decision-makers must understand how certain they can be of AI-driven predictions and account for potential uncertainty.

Balaprakash discussed ORNL's focus on developing multimodal AI systems that integrate various data sources, such as images, text, and sensor data, to generate more comprehensive insights. In the field of biology, for example, ORNL uses multimodal AI in plant phenotyping, where AI analyzes data from thermal imaging, RGB sensors, and other inputs to produce a holistic view of plant growth and structure. By tailoring models to the specific needs of different applications, ORNL balances accuracy, efficiency, and robustness, ensuring that AI solutions are both high-performing and suited to their operational environments.

Looking to the future, ORNL's Center for AI Security Research is leading efforts to secure sensitive data and model integrity through initiatives like secure enclaves, which create isolated environments for processing and protecting private information. In parallel, ORNL is advancing large foundational models across scientific disciplines, using unsupervised data that can later be refined for specific applications. By developing these foundational models, ORNL seeks to establish a scalable AI framework that reduces future development time and increases adaptability, allowing these large models to be customized for varied scientific and security-related challenges.

In his closing remarks, Balaprakash discussed how the workflows for AI are constantly evolving, and he emphasized the importance of iterative improvements to keep AI models aligned with changing requirements and new insights. During the following discussion, he elaborated on the trade-offs involved in training models, explaining that while a single, large model may be optimal for some applications, smaller models can often be more energy-efficient and easier to update. In applications requiring constant updates, smaller models provide flexibility without the extensive computational resources a single large model demands. These insights illustrate ORNL's dedication to advancing AI systems that perform robustly across various domains while adhering to stringent ethical and security standards.

Overall, Balaprakash's presentation exemplified ORNL's commitment to responsible AI. The lab's efforts underscore the delicate balance between technological progress and societal impact, showing that it is possible—and necessary—to innovate while remaining conscientious of the ethical, security, and operational demands that AI presents. ORNL's work in secure, efficient, and trustworthy AI for science and national security represents a promising path forward, where AI can empower researchers and policymakers to achieve their goals responsibly and sustainably.

Note: this Tech Note was written with the help of generative AI, based on a CI Compass Webinar given by Balaprakash on September 23, 2023, and reviewed by Balaprakash. The webinar can be viewed here: https://youtu.be/2wccATPl72A?feature=shared

This project is supported by the U.S. National Science Foundation Office of Advanced Cyberinfrastructure in the Directorate for Computer Information Science under Grant #2127548.