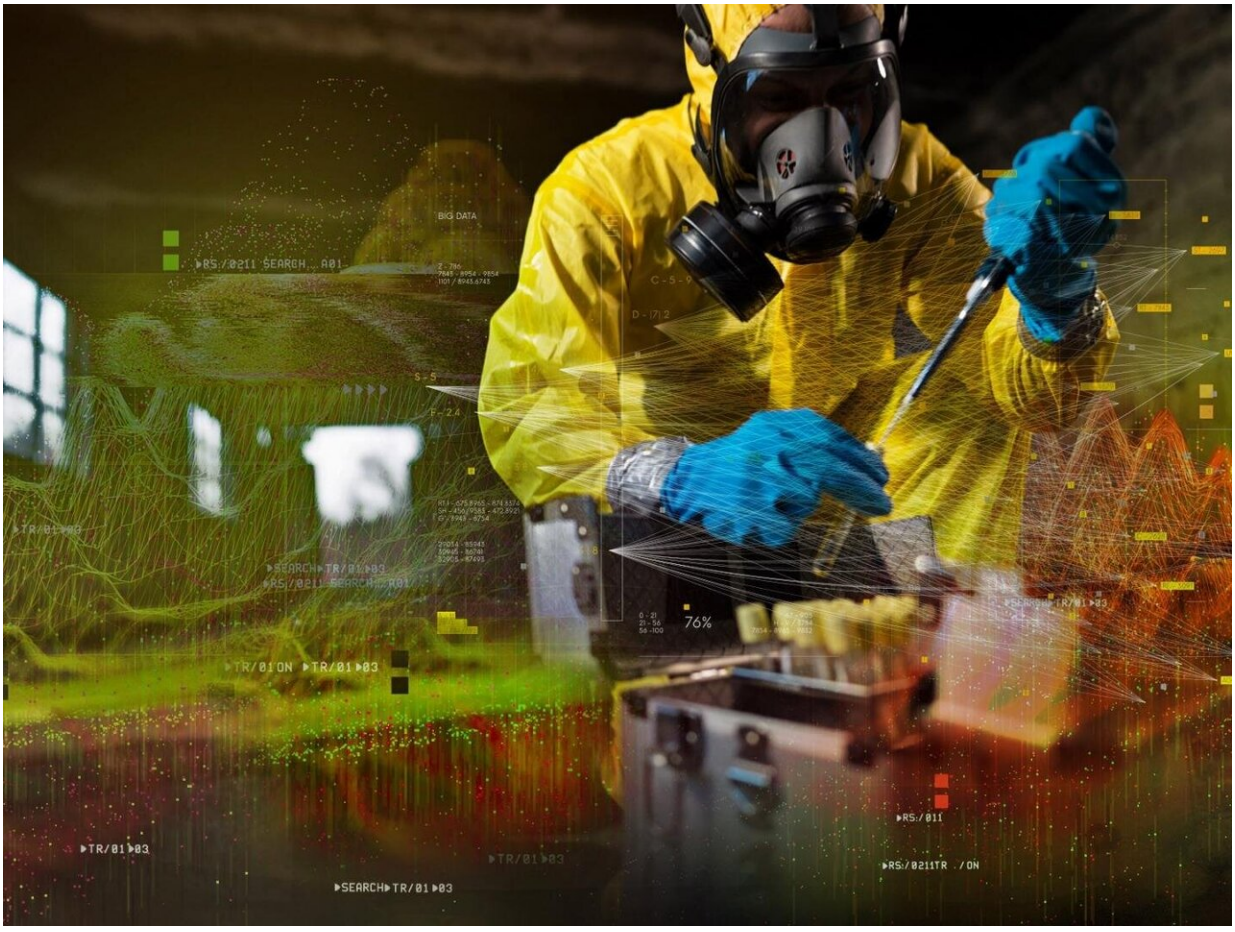
Explainable AI: A must for nuclear nonproliferation, national security

February 26 2021, by Tom Rickey



Explainable AI is enabling new ways to support long-term missions in national security, from mitigating biological and chemical threats to detecting and monitoring nuclear explosions around the globe. Credit: Timothy Holland / Pacific Northwest National Laboratory

We've all met people so smart and informed that we don't understand what they're talking about. The investment advisor discussing derivatives, the physician elaborating about B cells and T cells, the auto mechanic talking about today's computerized engines—we trust their decisions, even though we do not completely grasp the meaning of their words.

As it is with raw human intelligence, so it is with artificial intelligence (AI). We may not know exactly what's going on inside that elaborate black box built by humans, but its decisions can be so accurate that it earns our trust, if not our comprehension.

That's no problem when the decisions are of little consequence. Do we really need to understand, for instance, the inner workings of an AI system that sorts hundreds of photos of cats and dogs flawlessly in the time it takes to say the words "cats and dogs"?

Probably not. Human lives and the fate of nations do not rest on those decisions.

But the need for understanding escalates when the stakes are higher. For national security concerns under study at Pacific Northwest National Laboratory (PNNL), it's not good enough to know that a system works; scientists demand to know how and why.

Explainable AI: A path to understanding

That's the foundation for a field of study known as "explainable AI." The goal is to understand and explain the reasoning of the system—to untangle the threads of information that an AI system uses to make choices or recommendations.

"In the national security space, decisions are made by people who

demand transparency with the technologies they are working with," said Angie Sheffield, a senior program manager with the Department of Energy's National Nuclear Security Administration (NNSA).

Sheffield manages the data science portfolio in NNSA's Office of Defense Nuclear Nonproliferation Research and Development, also known as DNN R&D. The office oversees and improves the nation's ability to detect and monitor nuclear material production and movement, weapons development, and nuclear detonations across the globe. The office is supporting work by a team of PNNL scientists who are exploring how to make AI explainable in new ways.

AI is everywhere these days, from drug design to online purchasing, reservation systems, and health risk checklists. The national security realm, with vast data analysis challenges and powerful computing capabilities, is no exception. The stakes are exceedingly high when it comes to nuclear nonproliferation issues, and knowing exactly how an AI system reaches its conclusions is crucial.

"It can be difficult to incorporate a new and disruptive technology like AI into current scientific approaches. One approach is to build new ways that humans can work more effectively with AI," said Sheffield. "We must create tools that help developers understand how these sophisticated techniques work so that we can take full advantage of them."

Scarce data make explainable AI essential

The most common technique for training an AI system is presenting it with reams of data. When there are near-limitless photos of faces available, for example, a system learns a lot about the nuances involving eyes, noses, and mouths to use facial recognition to determine whether a glance from you should open your phone. The AI system relies on the

availability and input of a huge amount of data to enable the system to classify information correctly.

But—thankfully—data are much more sparse when it comes to nuclear explosions or weapons development. That good news complicates the challenge of using AI in the national security space, making AI training less reliable and amplifying the need to understand every step of the process.

"We're working to understand why systems give the answers they do," said Mark Greaves, a PNNL scientist involved with the research. "We can't directly use the same AI technologies that Amazon uses to decide that I am prepared to buy a lawn mower, to decide whether a nation is prepared to create a nuclear weapon. Amazon's available data are massive, and a mistaken lawn mower recommendation isn't a big problem.

"But if an AI system yields a mistaken probability about whether a nation possesses a nuclear weapon, that's a problem of a different scale entirely. So our system must at least produce explanations so that humans can check its conclusions and use their own expertise to correct for AI training gaps caused by the sparsity of data," Greaves added. "We are inspired by the huge advances that AI is continuing to make, and we are working to develop new and specialized AI techniques that can give the United States an additional window into nuclear proliferation activity."

A pinch of AI, a dash of domain knowledge

Sheffield notes that PNNL's strengths spring from two sources. One is significant experience in AI; PNNL scientists are frequent presenters at conferences that also feature researchers from entities such as Google, Microsoft, and Apple. But the other is domain knowledge—technical

details understood by staff at PNNL about issues such as how plutonium is processed, the type of signals unique to nuclear weapons development, and the ratios of isotopes produced by such materials.

The combination of data science, artificial intelligence, and national security experience gives PNNL a unique role in protecting the nation in the AI—national security space. It's combining the raw scientific power of AI with the no-nonsense street smarts of a nuclear sleuth.

"It takes a special set of knowledge, skills, and technical ability to advance the state of the art in national security," Sheffield said. "The consequences of what we do are very high, and we must go far beyond standard practice to be responsible."

Provided by Pacific Northwest National Laboratory