AI Readiness for the ePIC Detector

Dr. Markus Diefenthaler (Jefferson Lab)

Discussion Goals

1. Agree on a working definition of AI readiness in the context of the ePIC Detector.
2. Establish AI readiness considerations to be applied to each AI application.
3. Use the results to prioritize AI activities for the next three years.

This discussion is based on a memo on AI readiness that I prepared with my colleagues, Laura Biven and Torri Jeske.

AI Readiness for the ePIC Detector

Definition: **AI readiness** in the context of the ePIC detector refers to the extent to which the detector, its computing, data, and the associated workflows can support the development, integration, and sustained application of AI technologies.

Sustained application refers to the ongoing, routine, and reliable use of the AI technology as part of the system's operation over extended periods of time.

Example areas:

- Autonomous experimentation and detector control:
  - Fault detection, including anomaly detection,
  - Feedback loops between the DAQ-computing system and the detector subsystems.
- Fast decision-making in the orchestration of workflows such as alignment and calibration, monitoring, or reconstruction.
- Re-interpretable data through re-tunable workflows of data processing and multi-modal data.

Considerations for Each AI Application

1. The availability of **high-fidelity simulated datasets** for training and validation of AI models.
2. **Experimental data** (intended in a broad sense to include data from the detector-computing system and associated data streams such as slow control data and logbooks) for AI inference and validation.
3. Availability of **inference engines**, deployable at the edge (ASICS and DAM/FPGA boards) and within the computing ecosystem (CPUs, GPUs) to achieve latency requirements.
4. **Uncertainty quantification and interpretability**: Guardrails and fallback policies when models are uncertain and/or inputs are OOD.
5. **MLOps**: Sustainability of AI models relative to data drift, out of distribution data, etc. Clearly defined ownership of AI services, including responsibility for model development, deployment, monitoring, documentation, and incident response.
6. **Reproducibility** and **reusability** of scientific results through the capture of provenance, data, models, etc.; use of persistent identifiers; and support for long-term storage and sharing of data and code.
7. Alignment with standards (e.g. data models, data format, data ontologies…) to support **interoperability** of data (multi-modal) and systems.
8. **The human factor**: designing for humans in the loop in an attention-aware manner. By attention-aware, we mean in the sense that humans cannot physically monitor all things continuously. Systems can and should be designed around this fact to augment human capabilities.

AI Activity Priorities for the Next Three Years

1. The availability of **high-fidelity simulated datasets** for training and validation of AI models.
2. **Experimental data** (intended in a broad sense to include data from the detector-computing system and associated data streams such as slow control data and logbooks) for AI inference and validation.
3. Availability of **inference engines**, deployable at the edge (ASICS and DAM/FPGA boards) and within the computing ecosystem (CPUs, GPUs) to achieve latency requirements.
4. **Uncertainty quantification and interpretability**: Guardrails and fallback policies when models are uncertain and/or inputs are OOD.
5. **MLOps**: Sustainability of AI models relative to data drift, out of distribution data, etc. Clearly defined ownership of AI services, including responsibility for model development, deployment, monitoring, documentation, and incident response.
6. **Reproducibility** and **reusability** of scientific results through the capture of provenance, data, models, etc.; use of persistent identifiers; and support for long-term storage and sharing of data and code.
7. Alignment with standards (e.g. data models, data format, data ontologies…) to support **interoperability** of data (multi-modal) and systems.
8. **The human factor**: designing for humans in the loop in an attention-aware manner. By attention-aware, we mean in the sense that humans cannot physically monitor all things continuously. Systems can and should be designed around this fact to augment human capabilities.