UNITED STATES
NUCLEAR REGULATORY COMMISSION
ADVISORY COMMITTEE ON REACTOR SAFEGUARDS
WASHINGTON, DC 20555 - 0001

December 18, 2025

The Honorable David A. Wright
Chairman
U.S. Nuclear Regulatory Commission
Washington, D.C. 20555-0001

SUBJECT: SUMMARY REPORT – 731st MEETING OF THE ADVISORY COMMITTEE ON REACTOR SAFEGUARDS, DECEMBER 3 THROUGH 4, 2025

Dear Chairman Wright:

During its 731st meeting held December 3 through 4, 2025, which was conducted in person and virtually, the Advisory Committee on Reactor Safeguards (ACRS) discussed several matters. The Committee's efforts continued to be focused on topics highlighted in Executive Order (EO) 14300, "Ordering the Reform of the Nuclear Regulatory Commission." During this meeting the ACRS completed the following correspondence:

MEMORANDA

Memoranda to Mike King, Acting Executive Director for Operations, U.S. NRC, from Marissa G. Bailey, Executive Director, ACRS:

- December 2025 ACRS Full Committee – Topical Reports, dated December 10, 2025, Agencywide Documents Access and Management System (ADAMS) Accession No. ML25343A093,

- Documentation of Receipt of Applicable Official Nuclear Regulatory Commission (NRC) Notices to the ACRS for December 2025, dated December 10, 2025, ADAMS Accession No. ML25343A095, and

- Regulatory Guide, dated December 10, 2025, ADAMS Accession No. ML25343A090.

HIGHLIGHTS OF KEY ISSUES

In accordance with the Federal Advisory Committees Act, the agenda for this meeting was published in the *Federal Register* on November 18, 2025.

A. Palisades Steam Generator Operational Assessment (Restart Activities)

The Committee had an informational briefing from NRC staff during this meeting on the Palisades Nuclear Plant Restart activities. Specifically, the staff discussed a recent Steam Generator (SG) Operational Assessment that the licensee had submitted on the docket, and

the NRC staff reviewed. This is a time-tested, robust and extensively used industry SG monitoring and assessment methodology representing many reactor-years of operations experience. Discussions on margins, modeling, flaw growth projections, mitigative and corrective measures were held. The Committee also heard from several members of the public, documented in the Full Committee meeting transcript, and received written correspondence from members of the pubic (ADAMS Accession No. ML25342A161) which is attached to the meeting transcript file. The Committee members did not identify any new safety concerns and reiterated the Committee's conclusion from the September 22, 2025, letter report that NRC inspection staff should maintain a heightened vigilance in oversight during the first operating cycle and any subsequent cycles with the existing SGs. In addition, any signs of elevated primary-to-secondary leakage need to be highly scrutinized during those operational cycles. Finally, the ACRS support staff forwarded the correspondence and verbal comments received by the Committee to the Office of Nuclear Reactor Regulation Palisades Restart team.

B. Self-Assessment, Lessons Learned and Discussions during the Planning and Procedures

The ACRS Chairman highlighted that the topic identified in the agenda as self-assessment and lessons learned is not yet mature enough to warrant deliberation by the full committee (FC) and will likely be taken up during the February 2026 meeting planned for February 5 through 6. The Committee then moved into discussions associated with planning and procedures (P&P).

1. The Committee discussed the FC and subcommittee (SC) schedules through May 2026 as well as the planned agenda items for FC meetings.

2. The Committee discussed the recommendation to not review Regulatory Guide (RG) 1.220, Revision 0, "Acceptability of [American Society of Mechanical Engineers] ASME OM-2 Code, Component Testing Requirements at Nuclear Facilities, for New and Advanced Reactors." The Committee agreed and documented its decision in a memorandum dated December 10, 2025, ADAMS Accession No. ML25343A090.

3. The Committee heard from the SC Chairman regarding a SC meeting that was conducted prior to the government shutdown and since the November 2025 FC meeting, to include:

   - Accident Analysis SC meeting on Westinghouse topical report (TR) on Full Spectrum Loss-of-Coolant-Accident (LOCA) (September 16, 2025) [Member Martin]. The Committee documented its review of this document later in this report under item B.7.

4. The Committee discussed the following ongoing and near-term new reactor application reviews:

   a. Long Mott Generating Station (LMGS) construction permit application (CPA). Member Martin led this discussion and mentioned that he prepared a draft report on the unique, novel, and noteworthy (UNN) aspects of this application, for deployment of an X-energy Xe-100 nuclear power plant. This draft builds upon the earlier preliminary assessment (reported to the committee at the November 2025 P&P meeting) by organizing UNN elements around five Xe-100 Required Safety Functions (RSF) and two cross-cutting methodological issues. The RSF-based focus

areas are: 1) retain radionuclides in fuel particles and pebble graphite; 2) control of heat generation/reactivity; 3) control of heat removal; 4) control of water/steam ingress; and 5) maintain geometry. The sixth focus area addresses the sufficiency of the probabilistic risk assessment (PRA) centered licensing modernization project framework safety case. The seventh addresses the use of the Initial Test Program (ITP) to provide commissioning-stage confirmation of passive safety features. The current draft has been shared with committee members for internal review. The report concludes that the highest-priority elements for Committee focus are: (1) the scope and completeness of the PRA and hazard coverage supporting the construction-permit safety case; (2) the identification of cliff-edge behaviors and the adequacy of defense-in-depth for unanalyzed safety related and non-safety related with special treatment failure modes; (3) the first full demonstration of functional containment supported by a mechanistic source term traced from fuel performance through emergency planning; and (4) the trip-independent passive-safety demonstrations to be provided in the ITP.

NRC staff have independently prepared and provided ACRS with their list of UNN items for the LMGS CPA, prioritizing PRA scope, mechanistic source term, materials qualification, and selected ITP tests. Their list, while narrower in scope, is generally consistent with Member Martin's assessment. Collectively, the two perspectives are complementary, with the Committee’s draft providing a more complete basis for framing ACRS review priorities.

The CPA was submitted in May 2025, and the most recent estimate places the ACRS final review of the site-specific CPA in the August/September 2026 timeframe. Among the vendor-level Xe-100 topical activities expected ahead of the CPA review, only the Fuel Qualification TR is likely to come before the Committee, and it has not yet been scheduled.

The staff plans to issue the advanced safety evaluation to ACRS scheduled for August 2026. A letter report soon after is planned to support issuance of the final safety evaluation.

It was agreed that members would provide input on the UNN list by January 1, 2026.

b. Clinch River CPA application. Member Harrington led a discussion of this activity. He stated that he drafted an initial assessment of the UNN features included in the CPA to assist in defining the scope, level of detail, and timing of subcommittee review meetings. The first draft of a UNN summary paper was distributed to the members on November 11, 2025, for review and any feedback and comments are appreciated.

As an advanced boiling water reactor (BWR), the potential UNN features are much more narrowly focused than for non-light-water reactor designs coming before the committee. Nevertheless, as a passive design there are aspects that deviate from typical legacy BWR designs or may have limited or no prior demonstration in operational units and thus may rise to the level of UNN. These include the reactor vessel height to drive natural circulation, use of reactor isolation valves close-coupled to the vessel and their interaction with the operation of the isolation condenser system, no reactor coolant pressure boundary safety and relief valves, steel composite containment design, passive containment cooling system, and only

direct current power for design basis accident mitigation. More notably though is the use of the uniquely defined Safety Strategy in developing and defending the safety case and in implementing the safety classification of structure, systems and components (SSCs) for the design. Finally, areas in the preliminary safety analysis report where an alternate approach to regulatory guides or other regulatory precedent is identified have also been reviewed for UNN implications.

Member Harrington recommended that the Committee begin its review with the Safety Strategy TR. While it is not a radical departure from past practice in safety case development, it is nonetheless a new approach that warrants a close look and provides necessary context for evaluating the results.

For scheduling purposes, a final letter will likely be needed as a product of the July 2026 FC meeting.

It was agreed that members would provide input on the UNN list by January 1, 2026.

c. The Committee discussed briefly the following other ongoing and near-term new reactor applications: Atomic Alchemy construction permit application and the Fermi America combined license application for four AP1000s in Texas.

5. The Committee discussed the status of high priority rulemaking activities including:

   a. Title 10 of the Code of *Federal Regulations* (10 CFR) Part 57 (microreactors and low consequence reactors). Member Petti reiterated that the Committee plans to have a FC meeting on this topic and write a letter (currently planned for March 2026) after the rule is published for public comment in the *Federal Register*.

   b. The EO 14300 Rulemakings. Member Petti also led a discussion on this topic. Subcommittee Engagements are being scheduled for Section 5 of EO 14300. The Committee will decide which areas to review in Sections 5(b), 5(d), 5(e), 5(f), 5(g) for reactor security rules, and 5(h). If necessary, letters will be scheduled when the draft proposed rules are publicly available in the Spring of 2026. Letter reports to be scheduled, as needed, and after the draft rule is published for public comment in the *Federal Register*.

6. The ACRS Executive Director led a discussion of the following TR review recommendations and documented the Committee's decision in a memorandum dated December 10, 2025. These support new reactor applications.

   a. OKLO TR on Principal Design Criteria (PDC) for the AURORA Powerhouse. Chairman Kirchner led a discussion on this TR.

      The TR provides information on the development of PDC for the OKLO Aurora Powerhouse, based on RG 1.232, "Guidance for Developing Principal Design Criteria for Non-Light Water Reactors." The TR references an "Aurora powerhouse," but only uses that terminology to describe the nuclear power plant generically (pool type sodium fast reactor (SFR) technology) and does not refer to specific SSCs. Reference is made to the overall safety of the Aurora reactor design being based on inherent features and passive safety functions typical of SFRs, as demonstrated at the Department of Energy's EBR-II reactor and Fast Flux Test Facility. Two specific design features cited include metal fuel and a passive, always-on reactor vessel auxiliary cooling system. Power level is not cited in the TR, but available information suggests a 250 Megawatts thermal/75 Megawatts electric reactor module is contemplated.

In developing their PDCs, OKLO followed guidance in RG 1.232 by first assessing the applicability of SFR design criteria (SFR-DC), whether modification was required to specific design features of the Aurora powerhouse, and prior to modification of SFR-DC, whether advanced reactor design criteria or modular high temperature gas cooled reactor design criteria (MHTGR-DC) could be directly adopted. Notably, OKLO has taken the approach of adopting functional containment performance criteria (SECY-18-0096) and replaced SFR-DC 16, 38-43, and 50-57 with MHTGR-DC 16, 71, and 72. The Aurora powerhouse does not rely on electrical power to support any important to safety functions during anticipated operational occurrences or postulated accidents, deleting SFR-DC 17 and 18. SFR-DC 19 is modified to reflect that "during power operation, the plant is maintained in safe and stable conditions by automatic controls, and operator action is not required." The justification states, "The Aurora powerhouse does not have a traditional control room and instead utilizes an onsite monitoring room," which may be considered UNN. In addition, OKLO would delete SFR-DC 70 and 75-77 (intermediate cooling system integrity) based on the "failure of the intermediate coolant system has no significant impact on the safety of the plant," which may also be considered UNN.

Chairman Kirchner mentioned the need to obtain more information about the specific design being referenced in the TR. It was agreed that the ACRS staff would arrange an informal discussion with the NRC staff to better understand the design information before a decision is made to review the TR.

b. Electric Power Research Institute (EPRI) TR 3002032184, "U.S. Industry Performance Monitoring Inspection Plan for Select ASME Code Examination Items of [Pressurized Water Reactor] PWR Steam Generators and Pressurizers." Member Sunseri led a discussion on this TR.

This report is a continuation of the research in EPRI Reports 3002014590, 3002015906, and 3002023713 for SGs and Report 3002015905 for pressurizers (PZRs), which concluded that the time between inspections of SG and PZR components can be increased without compromising plant safety. Plants have used these reports as the technical basis for seeking and obtaining optimized inspection plans from the NRC from current American Society of Mechanical Engineers Boiler and Pressure Vessel Code, Section XI inspection requirements. Rather than plants seeking optimized inspection plans individually from the NRC, this report provides a comprehensive inspection plan for all SGs and PZRs in the U.S. fleet. It incorporates the NRC's requirement for adequate preventive maintenance (PM) to ensure that the occurrence of a novel degradation mechanism is identified in a timely manner. Key findings from the TR were:

- A comprehensive fleet-wide optimized inspection plan for SGs and PZRs in the U.S. PWR fleet has been developed to provide relief from the current ASME Code, Section XI inspection requirements without compromising safety.

- The plan provides adequate PM to meet the 25% sampling criterion implicit in the NRC's binomial distribution model.

- The plan includes recognition of units with existing PM commitments obtained through "Request for Alternative" submittals and subsequent NRC safety evaluation reports (SERs).

- Due to the dynamic and fluid nature of the inspection plan, it will be updated on a regular basis.

Member Sunseri has reviewed the report and determined that there are no unique, novel or noteworthy items contained in the report. Therefore, Member Sunseri recommends that no ACRS review be performed. The Committee agreed with the recommendation.

c. Framatome TR, ANP-10323, Revision 1, Supplement 1P, Revision 0, "One GALILEO: Fuel Rod Thermal-Mechanical Fuel Rod Methodology for PWRs." Member Palmtag led a discussion on this topic.

This TR describes four application methodologies to use the GALILEO fuel performance code to determine rod internal pressure, cladding oxide thickness, fuel centerline melt limits, and transient cladding strain limits. The methodologies do not change any approved GALILEO code models; the application methodologies use models that are already approved.

The methodologies appear to be very similar to the methodologies used with previous Framatome fuel performance codes, there does not appear to be anything unique, novel, or noteworthy in the TR. Member Palmtag recommended no ACRS review. The Committee agreed with the recommendation.

d. Framatome TR ANP-10300 Revision 1, Supplement 1P (Revision 0), "AURORA-B: Addressing [Limitations and Conditions] L&Cs 11, 12, and 18." Member Martin led a discussion on this topic.

Framatome submitted TR ANP-10300 Revision 1, Supplement 1P (Revision 0), "AURORA-B: Addressing L&Cs 11, 12, and 18," in June 2025 (ADAMS Accession No. ML25167A306) to provide generic methods for resolving the NRC staff's L&C numbers 11, 12, and 18 described their previous Safety Evaluation of the AURORA-B evaluation model (ADAMS Accession No. ML17346B057). These items currently require plant-specific justification for uncertainty treatment (L&C 11), enhanced modeling features (L&C 12), and certain conservative measures versus full statistical evaluations (L&C 18). This TR aims to standardize these elements to reduce future submittals.

AURORA-B was approved in 2018 to support BWR transient and accident analyses. This supplement introduces minor departures from that original evaluation model. Specifically, it provides procedural guidance for statistical treatment, optional modeling features, and validation of conservative assumptions. While the ACRS previously reviewed the AURORA-B methodology, the methods described in this supplement are not significantly different from established industry practice.

This TR does not introduce new phenomena modeling, alter safety conclusions, or affect the applicability of previously approved analyses. The revisions presented principally improve the applicant's implementation efficiency. Consistent with the Committee's review emphasis on unique, novel, or noteworthy safety implications, this TR supplement does not present issues that meet the threshold for ACRS review. Member Martin recommended no Committee review of this TR. The Committee agreed with the recommendation.

7. Member Martin led a discussion of the documentation of the SC review of the Westinghouse TR, WCAP-18850, Adaptation of the FULL SPECTRUM LOCA (FSLOCA) Evaluation Methodology to Perform Analysis of Cladding Rupture for High Burnup Fuel.

   The ACRS Accident Analysis Subcommittee met on September 16, 2025, and reviewed this TR. The TR was built on prior approved WEC methods, FSLOCA Methodology (WCAP-16996-A, Revision 1) and Incremental Burnup Extension (WCAP-18446-A) for the analysis of their fuel products with AXIOM cladding with standard UO_2 or ADOPT pellets. The TR discussed that WCOBRA/ TRAC-TF2 code was modified to analyze higher burnup fuel with higher initial enrichment. It considers all higher burnup fuel rods in the core, and the analysis is focused on cladding rupture. Models were assessed and/or updated primarily to ensure that all fuel rod phenomena associated with higher burnup levels were appropriately captured to support high probability, licensing basis calculations. Methodology uncertainties and limitations and conditions were also discussed.

   The methodology addresses the primary technical challenge of fuel fragmentation, relocation, and dispersal (FFRD), which emerges at high burnup levels (above approximately 55-60 GWd/MTU). Westinghouse's approach avoids modeling fuel dispersal by demonstrating through a non-parametric statistical framework that cladding rupture does not occur with high probability for high-burnup fuel rods across three break spectrum regions (small, intermediate, and large breaks). By preventing rupture rather than modeling dispersed fuel behavior, the methodology provides a conservative and tractable approach to addressing FFRD concerns. The methodology requires parallel analyses. Specifically, these are a traditional emergency core cooling system compliance analysis per 10 CFR 50.46 using extended WCAP-16996 methodology and a rupture prevention demonstration using WCAP-18850.

   The SC found the methodology sound and noted its alignment with ongoing regulatory initiatives including the NRC's increased enrichment and burnup rulemaking, EPRI's Alternative Licensing Strategy approach, and the proposed performance-based framework in 10 CFR, Section 50.46c, positioning it to facilitate bundled licensing actions for power uprates, extended fuel cycles, and advanced fuel transitions. Member Martin recommended no Committee review of this TR. The Committee agreed with the recommendation.

8. Technical Support Branch Chief Burkhart led a discussion of the proposed 2027 meeting calendar which was distributed to all members prior to this meeting. The calendar was approved as proposed and will be placed on the ACRS SharePoint site.

9. There were no reconciliations during this meeting.

10. The Executive Director led a discussion of important notices including:

    - Conflict of Interest Guidance for Employees Participating in Duane Arnold Restart Activities
    - Summary of Conflict-of-Interest Prosecutions
    - Reminder on the Applicability of the Federal Rules of Ethics During Furlough Periods
    - Equal Employment Opportunity Policy Statement
    - Reminder: Please Update/Confirm Contact Information for the NRC Mass Notification System
    - Cooperating With and Providing Information to the Office of the Inspector General

11. Member Bier provided a researched presentation on artificial intelligence (AI) activities, which is attached to this summary report as Enclosure 2. The presentation noted that the rate of progress in AI technologies – particularly Generative Artificial Intelligence (AI), including Large Language Models and Retrieval Augmentation Generation is very rapid. Slide 3 of the presentation presents a summary of the key issues regarding AI and Slides 15 through 21 provides some conclusions and future recommendations for the Committee's consideration. The presentation is not an official ACRS position as the Committee only speaks through its letter reports. The research and slide deck is a knowledge management presentation developed by Member Bier and Consultant Myron Hecht for the benefit of ACRS Members at the December P&P meeting.

12. The Committee conducted its annual leadership elections in accordance with Section 8 of the bylaws. Newly elected leaders will assume their duties on January 1, 2026. The results of the elections were:

    Chairman: Gregory Halnon
    Vice Chairman: David Petti
    Member-at-Large: Craig Harrington

13. There was no closed session as part of this P&P.

14. The following topics are on the agenda for the 732nd ACRS FC meeting, which will be held February 5 through 6, 2026:

    - Self-assessment, lessons learned and path forward discussion.

Sincerely,

Walter L Kirchner Signed by Kirchner, Walter
on 12/18/25

Walter L. Kirchner
Chairman

Enclosures:
1. List of Acronyms
2. Notes on Artificial Intelligence Applications
   in Nuclear Power and Regulatory Implications

December 18, 2025

SUBJECT: SUMMARY REPORT – 731st MEETING OF THE ADVISORY COMMITTEE ON REACTOR SAFEGUARDS, DECEMBER 3 THROUGH 4, 2025

Accession No: ML25345A203 **Publicly Available (Y/N):** Y **Sensitive (Y/N):** N
If Sensitive, which category?
Viewing Rights: ☒ NRC Users or ☐ ACRS only or ☐ See restricted distribution

OFFICE	ACRS	SUNSI Review	ACRS	ACRS
NAME	LBurkhart	LBurkhart	RKrsek	WKirchner
DATE	12/11/2025	12/11/2025	12/17/2025	12/18/2025

OFFICIAL RECORD COPY

LIST OF ACRONYMS

10 CFR	Title 10 of the *Code of Federal Regulations*
ACRS	Advisory Committee on Reactor Safeguards
ADAMS	Agencywide Documents Access and Management System
AI	Artificial Intelligence
ASME	American Society of Mechanical Engineers
BWR	Boiling-water Reactor
CPA	Construction Permit Application
DBA	Design Basis Accident
EO	Executive Order
EPRI	Electric Power Research Institute
FC	Full Committee
FFRD	Fuel Fragmentation, redistribution, and dispersal
FSLOCA	Full Spectrum Loss of Coolant Accident
ITP	Initial Test Program
L&C	Limitation and Condition
LOCA	Loss-of-Coolant Accident
LMGS	Long Mott Generating Station
MHTGR-DC	Modular High Temperature Gas Reactor – Design Criteria
NRC	Nuclear Regulatory Commission
PDC	Principal Design Criteria
PM	Preventive Maintenance
PRA	Probabilistic Risk Assessment
PWR	Pressurized Water Reactor
PZR	Pressurizer
P&P	Planning and Procedures
RG	Regulatory Guide
RSF	Required Safety Functions
SC	Subcommittee
SFR	Sodium Fast Reactor
SFR-DC	SFR Design Criteria
SG	Steam Generator
TR	Topical Report
UNN	Unique, Novel, or Noteworthy

Notes on Artificial Intelligence Applications in Nuclear Power and Regulatory Implications

Myron Hecht and Vicki Bier

ACRS

Last Updated August 27, 2025

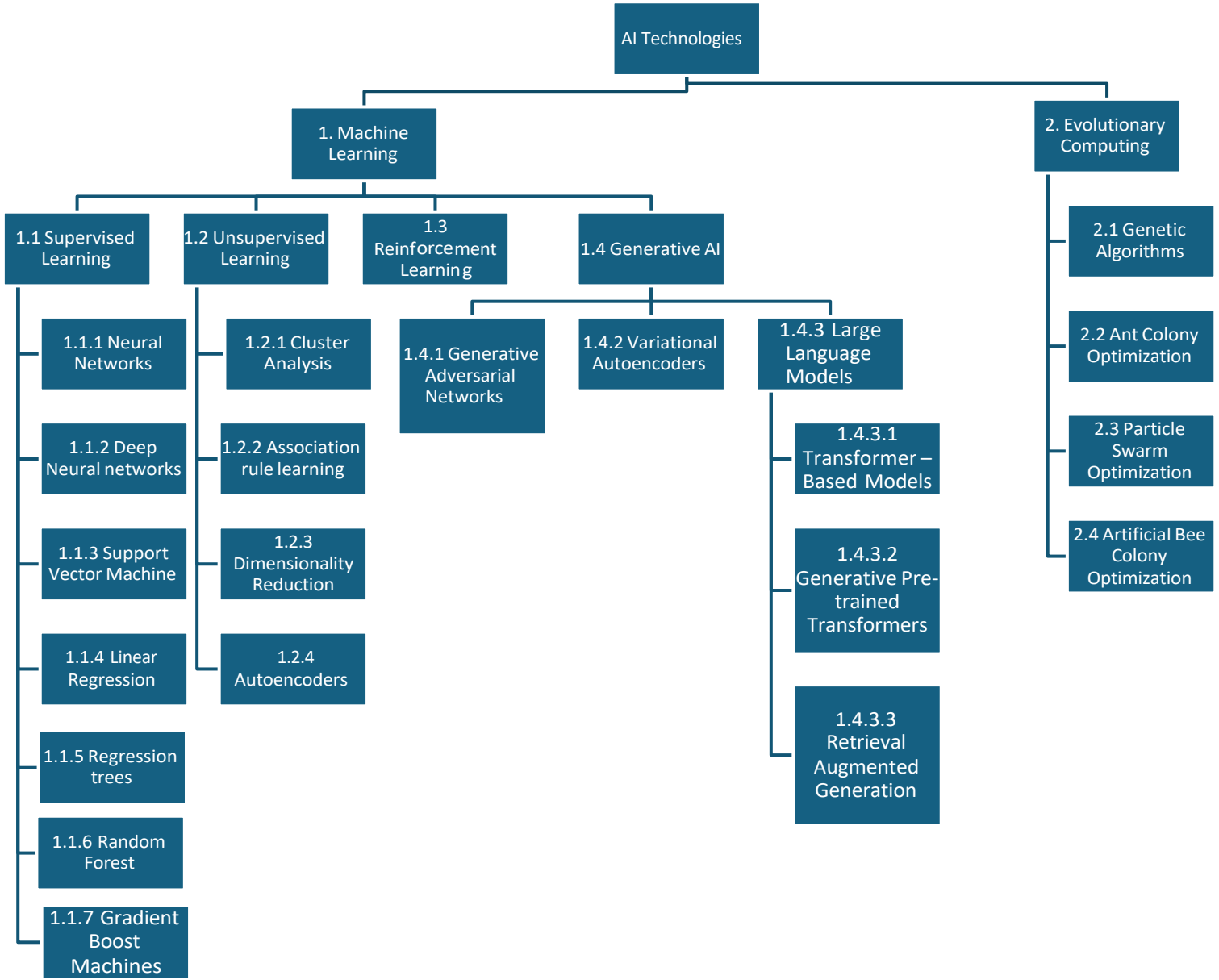
Enclosure 2

Contents

Summary

- **Difference between AI software and Traditional Software:** Unlike traditional software programmed to perform a task, AI software systems are programmed to learn to perform a task through a process called training (setting of coefficients called "weights" to minimize a difference between the input data and the output of the AI system)
- **Research on applications of AI to nuclear power:** Research in AI applications for nuclear power have been investigated since 1992 with use of a neural network for nuclear power plant status diagnostics. Other applications that have been investigated include Prognostics and Health Monitoring (PHM), real-time operator support, simulation, and nuclear power plant design.
- **The higher performing, the less trustworthy:** Higher performing AI algorithms such as deep neural networks lack explainability (the inability human users, stakeholders, and regulators their decisions, actions, and behaviors to trace and understand the reasons why AI systems reached their conclusions). This means that they are not as inherently trustworthy.
- **Causes of AI failures are different from conventional software:** AI failures are not merely the result of straightforward bugs or user errors. Instead, they emerge from interactions among algorithms, humans, and the specific situations in which they are deployed. The multifaceted nature of AI failures makes traditional verification and validation much less effective than it is for conventional deterministic software.
- **Most promising near-term applications of AI:** The most promising near-term application of AI is in PHM because of the level of research activity, availability of commercial products, and experience in applications in non-nuclear industries
- **Second most promising application:** Regulatory Queries and Compliance Monitoring using Retrieval Augmented Generation (RAG) .
- **Where AI software should never be used:** Closed loop real-time control software, with its deterministic outputs and testable response times, has been used for automated control of nuclear reactors for decades. However, because it is not possible to verify AI software containing both complex training algorithms and unverifiable training sets, AI-based real time control systems are not verifiable to the degree necessary for safety critical applications. Therefore, AI software should not be used in real-time reactor control or other applications with high consequences of control software failures.

Artificial Intelligence Taxonomy



See Backup for descriptions

AI Nuclear Applications Hazards, Causes, and Mitigations by Application Area (1/2)

Application Area	Description	Application	Hazards	Cause	Mitigation
Prognostics and health monitoring (Condition monitoring, fault diagnosis, predictive maintenance	Identify conditions of wear, estimate remaining useful life and identify anomalies using neural networks, multivariate adaptive regression, Support vector machines analysis from sensor data to	Transient diagnostics at the plant level [1], [2] Transient predictions for SMRs [52] Pressurizer transients [3] Fault detection of rotating bearings [4] Control rod drive mechanisms [5] Sensor failures [6,7,8] Batteries [9 ,10]	Failure to detect off-nominal conditions Spurious identification of off-nominal conditions leading to unnecessary maintenance actions Overestimate RUL Fail to detect anomalies	Insufficient training data, Biased, incomplete, or "poisoned" training data, errors in implementation of ML training and execution algorithms	1. Requirements to document training data 2. Development or adoption of methods to detect bias 3. Methods to detect adversarial training data 4. Verification of ML training algorithms 5. Verification of ML execution algorithms
Reactor control	Nuclear power plant operator support (not direct control)	Optimize power levels and flux distributions based on neural networks [21] Accounting for changing data such as demand, weather conditions [21] Optimizing based on equipment condition, fuel consumption [21]	Plant operators rely on real-time analysis applications using non-deterministic AI algorithms with limited or no explainability	Real time AI algorithms trained by incomplete, or "poisoned" training data, errors in implementation of ML training and execution algorithms Response time failures caused by non-deterministic algorithms	1. Requirements to document training data 2. Development or adoption of methods to detect bias 3. Methods to detect adversarial training data 4. Verification of ML training algorithms 5. Verification of ML execution algorithms 6. Watchdog deterministic algorithms to intervene and bring system to safe state

Table includes application selections from Huang, et. al. [21]

AI Nuclear Applications Hazards, Causes, and Mitigations by Application Area (2/2)

Application Area	Description	Application	Hazards	Cause	Mitigation
Simulators	Use of neural networks to simulate plant behavior in plant simulators	Training [20] Testing of new procedures [20] Digital Twins [11 ,12]	Non-deterministic algorithms may not accurately simulate plant behavior. Verification is complicated by non-explainable nature of deep neural networks nature of AI algorithms	Real time control AI algorithms trained by incomplete, or "poisoned" training data, errors in implementation of AI algorithms Algorithms that lack explainability Response time failures caused by non-deterministic algorithms	1. Requirements to document training data 2. Development or adoption of methods to detect bias 3. Methods to detect adversarial training data 4. Verification of ML training algorithms 5. Verification of ML execution algorithms 6. Watchdog deterministic algorithms to intervene and bring system to known state 7. Data recording to enable reconstruction of anomalous responses
Nuclear Power Plant Design	Thermal hydraulics Computation Fluid Dynamics Shielding Calculations	Deep Neural surrogate models of Computation Fluid Dynamics [13] and thermal hydraulics for computational acceleration [14] Back-propagation networks to determine heat transfer coefficients [15] AI-based methods for critical heat flux calculations [16, 17] Genetic algorithms for neutron shielding calculations [18] Neural network for dose estimation using weight, thickness, and materials in the input layer [19]	Errors in thermal hydraulics, computation fluid dynamics, and shielding design calculations leading to unsafe operating conditions, radiation exposure, or radioactive material release	Insufficient training data, Biased, incomplete, or "poisoned" training data, errors in implementation of ML training and execution algorithms	1. Requirements to document training data 2. Development or adoption of methods to detect bias 3. Methods to detect adversarial training data 4. Verification of ML training algorithms 5. Verification of ML execution algorithms
Regulatory Document summarization and queries	Compliance, Licensing and Enforcement	Combining and summarization of regulations from on a single topic from multiple sources using Retrieval Augmented Generation	Incomplete or incorrect aggregation of regulations	Defects in RAG algorithms, incomplete document set	Manual review of outputs

Observations on Nuclear Power Applications of AI

- Prognostics and Health Management (PHM) is an active area of research and the most promising area for initial application in Nuclear Power Plants [3, 4, 5, 6, 7, 8, 9, 10]
- Regulatory Queries and Compliance Monitoring using Retrieval Augmented Generation (RAG) is a very active of research in non-nuclear applications [48, 49, 50, 51] . We have not found any documentation of its application in the nuclear industry, but it is well suited to NRC regulatory activities and compliance (see description section 1.4.3.3 and charts following PHM)
- Publications on use of AI for Simulation and Nuclear Power Plant design found in this literature search are all experimental– no production use.
- Artificial Neural networks have been used for transient detection and prediction since 1992 [1, 2, 53] but not for actual control. Automated reactor control is done with traditional deterministic algorithms and software. There is not any compelling advantage to the use of AI for direct reactor control.
- Commercial products using AI for mechanical engineering design in non-nuclear applications are now available [28]. Whether any of these products are suitable for component or structure design in nuclear power applications is unclear. As was the case for reactor control, deterministic software applications has been and can be used.

AI for Prognostics and Health Management (PHM) (1/4)

- Non-AI PHM software systems are in use at nuclear power plants [34, 35, 36, 37]
- AI algorithms used in PHM (Support Vector Machines, Neural Networks, Generative adversarial networks) are an evolution of deterministic statistical analysis techniques in PHM systems (multiple regression, wavelet spectrum analysis, compressed sensing, multivariate signal processing, principal component analysis)
- AI applications in PHM is an active area of research [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 32, 33]
- Commercial PHM products incorporate AI algorithms (see following charts)

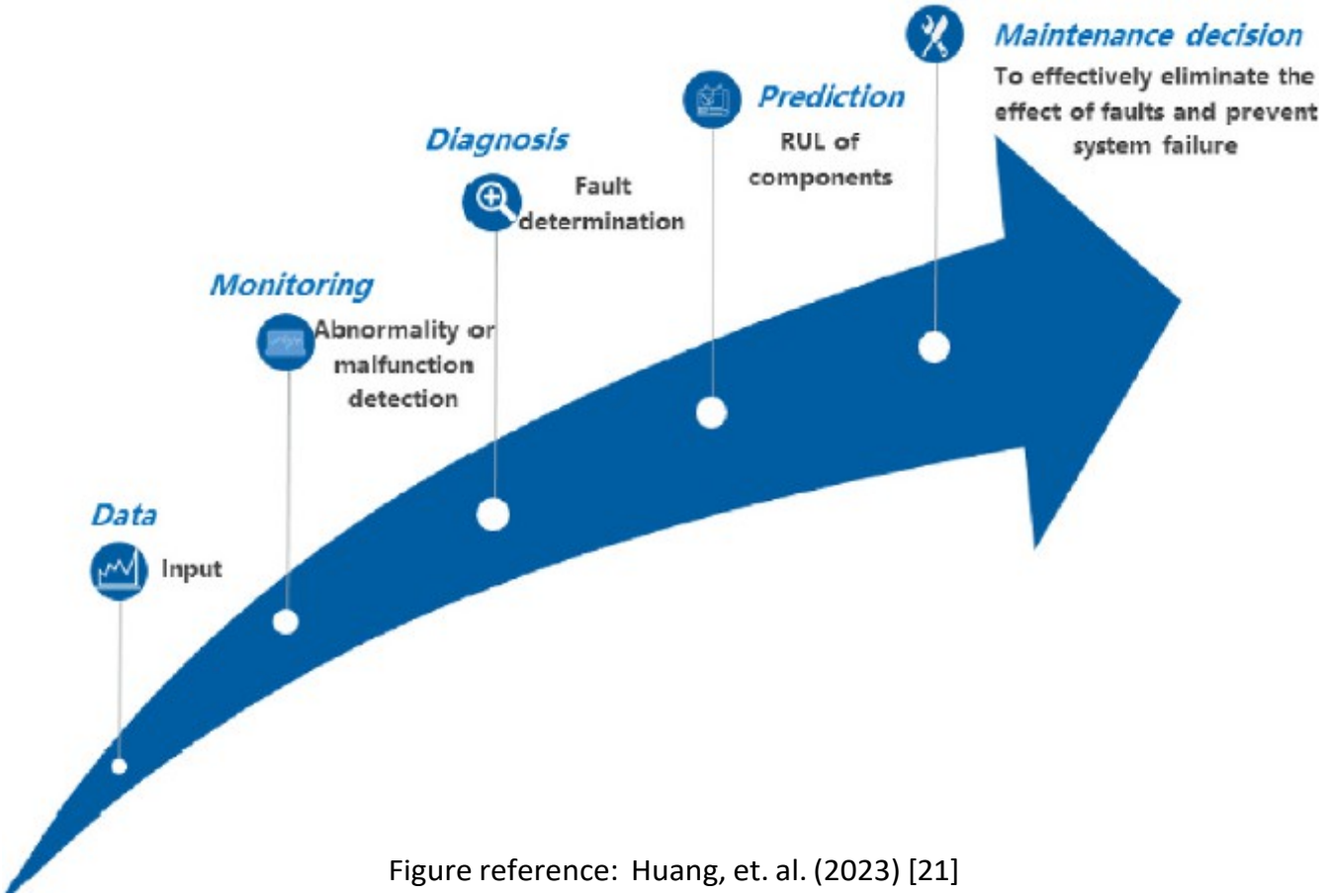


Figure reference: Huang, et. al. (2023) [21]

AI for Prognostics and Health Management (PHM) (2/4)

PHM Products Incorporating AI*

1. **Siemens MindSphere**: An industrial IoT platform that uses AI and advanced analytics to provide predictive maintenance and asset performance management [29]
2. **IBM Maximo**: An enterprise asset management solution that incorporates AI for predictive maintenance and asset health insights. [30]
3. **Uptake**: An AI-driven analytics platform that provides predictive insights for industrial asset management. [31]
4. **Honeywell Forge**: An enterprise performance management solution that uses AI to improve asset reliability and operational efficiency. [32]
5. **Siemens Senseye**: A predictive maintenance software that uses AI to provide prognostics and health management for manufacturing and industrial processes. [33]

*Contents of this chart were generated using a Large Language model with the prompt "What commercial software products for Prognostics and Health Management use AI?", Manually verified; three suspect results eliminated

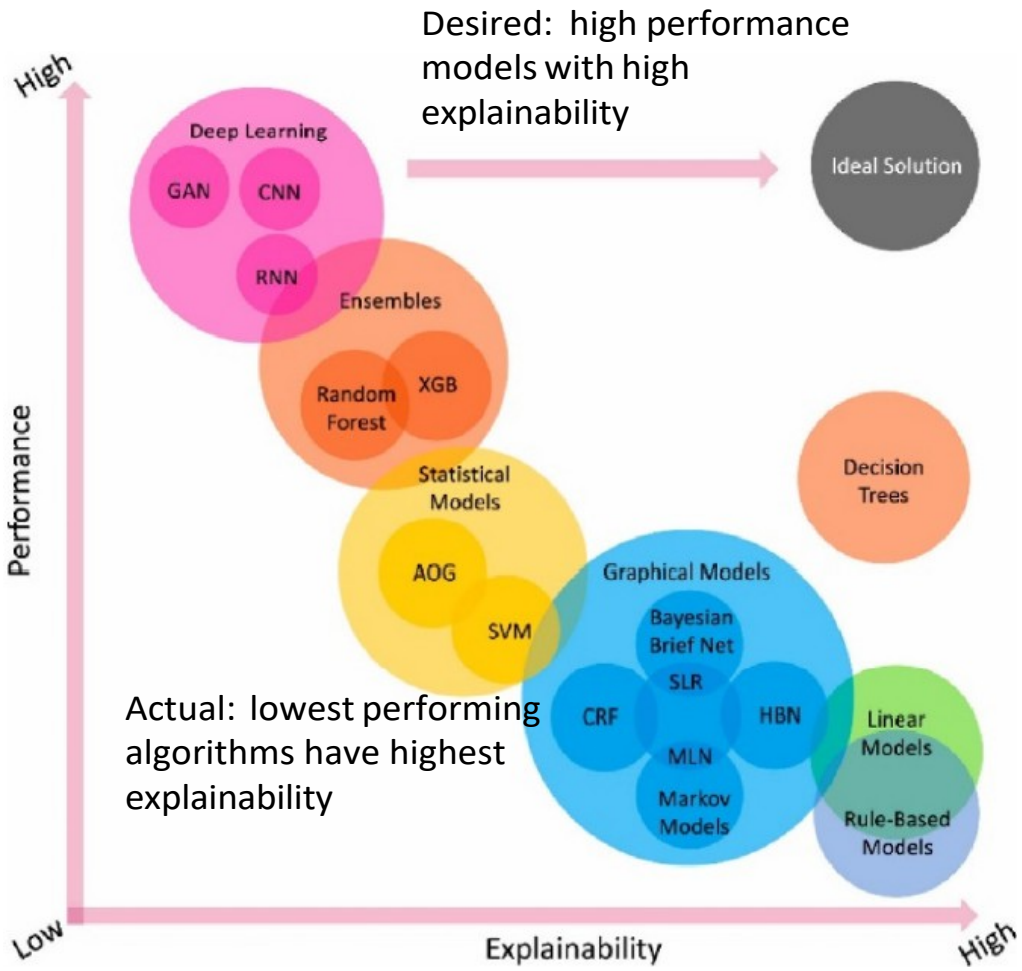
AI for Prognostics and Health Management (PHM) (3/4)

Concerns

- Limited training data: a high proportion of the data used to train and test models in the experimental PHM papers listed on the previous charts is derived from simulation calculations owing to the high-cost investment from the acquisition of experimental data of NPPs, as well as the risk of reproducing some complex operating conditions (such as severe accident scenarios). [7, 21] Can lead to unexpected responses
- Explainability: Models with better performance (highest PHM accuracy and lowest false positives) usually have a more complex internal mechanism and are therefore difficult to explain, whereas models with a clearer mechanism, such as linear models and rule-based models, lack the ability to deal with some complex nonlinear problems (see next chart) [21]
- Models with low explainability will be difficult or impossible to license.

AI for Prognostics and Health Management (PHM) (4/4)

The higher the performance of AI models, the worse their explainability*



- GAN: generative adversarial network.
- CNN: Convolutional Neural Network
- RNN: Recurrent Neural Network
- XGB: extreme gradient boosting;
- AOG: stochastic and–or graphs;
- SVM: Support Vector Machine
- HBN: hierarchical Bayesian networks
- SLR: simple linear regression;
- CRF: conditional random fields;
- MLN: Markov logic network;

*NIST Principles of Explainability

- **Explanation**: A system delivers or contains accompanying evidence or reason(s) for outputs and/or processes.
- **Meaningful**: A system provides explanations that are understandable to the intended consumer(s).
- **Explanation Accuracy**: An explanation correctly reflects the reason for generating the output and/or accurately reflects the system's process.
- **Knowledge Limits**: A system only operates under conditions for which it was designed and when it reaches sufficient confidence in its output

Figure reference: Huang, et. al. (2023) [21]

AI for regulatory document summarization and queries

- Retrieval Augmented Generation has been used to combine regulatory documents with a Large Language Model to create more accurate and up to date responses [48, 49, 50, 51]
- Combining agency specific documents with the data contained in a Large Language Model (LLM) provides accurate responses to queries using a wide range of documents (Regulations, Reg. Guides, Licensee material, SERs, NUREGs, endorsed industry standards, etc.) as an enhancement.
- Using agency specific documents together with the LLM internal provides greater specificity and reduces the likelihood hallucinations

AI Failure Mechanism Examples: Non-intentional

Mechanism Name	Description	Nuclear power example
Reward Hacking	Reinforcement Learning (RL) systems act in unintended ways or causes unintended side effects because the reward function does not account for all aspects of the situation or environment [26]	A robot in a radioactive environment spill knocks over an object in its path while performing a task because the reward function does not account for obstacle avoidance
Distributional shifts	The system is tested in one kind of environment, but is unable to adapt to changes in other kinds of environment [27]	A robot was able to avoid a radioactive spill successfully and reach its goal. During testing, they slightly moved the position of the radioactive spill, but the robot was not able to avoid it
Natural Adversarial Examples (also called Adversarial Learning)	[in neural networks, perturbations in the input data that cause an unintended alteration in the output, e.g., a pixel alteration that changes the perception of a red light to a green) [24, 27]	A surveillance camera system identifies a guard as an intruder
Common Corruption	The system is not able to handle common corruptions and perturbations such as tilting, zooming, or noisy images [26]	A noise in sensor data causes a PHM AI system to spuriously annunciate an anomaly .26
Insufficient testing	The ML system is not trained and tested in the full range of conditions in which it is intended to operate. For example, an autonomous vehicle at an intersection with a missing stop sign [25]	An AI assistant designing a pressurizer fails to include a relief valve

AI Failure Mechanism Examples: intentional

Name and Description of Attack	Nuclear Power Plant Example	Documentation of Attack
Perturbation attack: attacker stealthily modifies the input to get the desired result	Image classification system in nuclear power plant supplier fails to identify a defective item Transcription system recording an I&E meeting inserts wrong items in transcript	Image Classification: Noise is added to an X-ray image, which makes the predictions go from normal scan to abnormal [35] Text translation: Specific characters are manipulated to result in incorrect translation. The attack can suppress specific word or can even remove the word completely [36] Speech: Researchers showed how given a speech waveform, another waveform can be exactly replicated but transcribes into a totally different text. [37]
Poisoning Attacks: Attacker contaminates the model generated in training so that predictions no new data will be modified	Reactor control system increases power during event that should cause a shutdown	In a medical dataset where the goal is to predict the dosage of anticoagulant drug Warfarin using demographic information, etc. Researchers introduced malicious samples at 8% poisoning rate, which changed dosage by 75.06% for half of patients [38]
Reprogramming deep neural nets - an attacker writes a query to reprogram a task that deviates from the original intent	Image processing system in radwaste facility fails to identify unauthorized persons	Demonstrated how ImageNet, a system used to classify one of several categories of images was repurposed to count squares. [39]
Adversarial Example: An adversarial example is an input/query from a malicious entity sent with the sole aim of misleading the machine learning system. These examples can manifest in the physical domain.	Image processing in nuclear power plant receiving inspection system fails to identify defective items	Researcher creates a rifle with a 3D printer with custom texture that fools image recognition system into thinking it is a turtle [40] Researchers construct a sunglass with a design that can now fool image recognition system, and no longer recognize the faces correctly [41]
Backdoor Machine Learning - A trusted party is misusing their access in order to cause harm	PHM system is sabotaged so that fault diagnostics can not be trusted	Researchers created a backdoored U.S. street sign classifier that identifies stop signs as speed limits only when a special sticker is added to the stop sign (backdoor trigger) [42]
Exploit software dependencies of ML system. Attacker exploits the software vulnerabilities such as a buffer overflow within dependencies of the system.	Image processing system in radwaste facility fails to identify unauthorized persons	An adversary sends in corrupt input to an image recognition system that causes it to misclassify by exploiting a software bug in one of the dependencies.[43]

Recommendations

- The most promising near-term applications of AI in nuclear power
  - Prognostics and Heath Monitoring (PHM) because of the level of research activity and experience in applications in non-nuclear industries
  - Retrieval Augmented Generation (RAG) for the queries of regulatory and licensee documents for licensing evaluations, inspections, reviews and policy making
- Further research and regulations for AI deployment should address at least the following
  1. Documentation of training data
  2. Methods to assess the sufficiency and relevance of training data [55]
  3. Methods to assess the presence or absence of training data bias [53]
  4. Methods to detect adversarial training data [56]
  5. Verification of ML training algorithms
  6. Verification of ML execution algorithms
  7. Configuration management and methods of safeguarding training data and AI software from tampering [54]
  8. Quality and accuracy requirements from Large Language Models and RAG systems
  9. Explainability of AI System Results [56]
- Establish/maintain contacts with authoritative organizations
  - NIST Trustworthy and Responsible AI Center
  - IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems

Limitations of this Work

- The rate of progress in AI technologies – particularly Generative AI (including Large Language Models and Retrieval Augmentation Generation) is very rapid -- the descriptions and conclusions reached in this work may be superseded within months of its completion (May, 2025)
- Our search for nuclear applications of AI was a best effort, but we have no method for assessing coverage or completeness
- Assessment of failure mechanisms, failure modes, and hazards for nuclear power applications are examples, not complete lists

AI and Nuclear Safety

- AI could well contribute to enhanced nuclear safety, especially if designed and implemented as part of an integrated safety system
- This is likely to involve purpose-built AI tools:
  - Rather than generic off-the-shelf tools like large language models

  And tools designed with a focus on safety more so than efficiency
- We may need to “go slow to go fast”:
  - The nuclear industry can’t afford to “move fast and break things”

Caveats with AI for Nuclear Safety

- Complacency
- Need for human intervention in unanticipated situations
- Economic pressure to achieve greater efficiencies

Complacency

- AI can be useful as a supplement to human judgment
- For example, it's been shown that a radiologist plus an AI model outperforms two radiologists in interpreting mammograms:
  - The two radiologists likely share similar training and perspectives/biases
  - By contrast, the AI sees things in a completely different manner
- However, radiologists accustomed to relying on AI may no longer take the time to provide their own independent judgments:
  - In which case the expectations for level of accuracy of the AI would become much greater
- We may need a new term for "social loafing" when the other part of the team is an AI model

Need for Human Interventions in Emergencies

- Even the most accurate AI models are vulnerable when circumstances differ from what was in their training data ("distribution shift")
- "Label shift":
  - The prevalence of different circumstances may change
  - E.g., hospitals in white-collar vs. blue-collar areas, or U.S. vs. India
- "Concept shift":
  - New phenomena may emerge
  - E.g., the advent of COVID pneumonia changed interpretation of chest X-rays
- How does this apply to nuclear safety?
  - E.g., an earthquake may change the interpretation of plant monitoring data
  - Same applies to operation outside the safe operating envelope (ATHEANA)

Economic Pressure to Achieve Efficiencies

- As noted, AI can be useful as a supplement to human judgment:
  - E.g., NRC inspectors using AI reviews of plant data to set inspection priorities
  - This can lead to better-focused inspections, with less preparation time
- However, there are always efficiency pressures:
  - E.g., inspectors may be expected to do twice as many inspections per year
  - This will "eat up" the improvement in inspection quality, and encourage complacency (over-reliance on AI results)
- The same thing may happen in AI-aided or semi-autonomous operation:
  - E.g., AI may make it possible for one operator to oversee multiple reactors
  - However, there may be pressure to increase the number of those reactors over time, until it eventually becomes excessive, causing or exacerbating accidents

References

[1] E.B. Bartlett and R.E. Uhrig, Nuclear power plant status diagnostics using an artificial neural network, Nucl. Technol. 97 (3) (1992) 272–281.

[2] A. Basu, E.B. Bartlett, Detecting faults in a nuclear-power-plant by using dynamic node architecture artificial neural networks, Nucl. Sci. Eng. 116 (4) (1994) 313–325.

[3] P. Baraldi, F. Di Maio, E. Zio, Unsupervised clustering for fault diagnosis in nuclear power plant components, Int. J. Comput. Intell. Syst. 6 (4) (2013) 764–777.

[4] D. Miki, K. Demachi, Bearing fault diagnosis using weakly supervised long short-term memory, J. Nucl. Sci. Technol. 57 (9) (2020) 1091–1100.

[5] A. Oluwasegun, J.C. Jung, The application of machine learning for the prognostics and health management of control element drive system, Nucl. Eng.

[6] J. Choi, S.J. Lee, A sensor fault-tolerant accident diagnosis system, Sensors 20 (20) (2020).

[7] S.M. Zhu, et al., A robust strategy for sensor fault detection in nuclear power plants based on principal component analysis, Ann. Nucl. Energy (2021) 164.

[8] Y. Yu, et al., Improved PCA model for multiple fault detection, isolation and reconstruction of sensors in nuclear power plant, Ann. Nucl. Energy (2020) 148.

[9] F.S. Lasheras, et al., A hybrid PCA-CART-MARS-based prognostic approach of the remaining useful life for aircraft engines, Sensors 15 (3) (2015) 7062–7083.

[10] M.A. Patil, et al., A novel multistage Support Vector Machine based approach for Li ion battery remaining useful life estimation, Appl. Energy 159 (2015) 285–297.

[11] M.I. Radaideh, et al., Neural-based time series forecasting of loss of coolant accidents in nuclear power plants, Expert Syst. Appl. (2020) 160.

[12] B. Kochunas, X. Huan, Digital twin concepts with uncertainty for nuclear power applications, Energies 14 (14) (2021).

[13] Q. Lu, et al., Prediction method for thermal-hydraulic parameters of nuclear reactor system based on deep learning algorithm, Appl. Therm. Eng. (2021) 196.

[14] K. Cheng, et al., Development and validation of a thermal hydraulic transient analysis code for offshore floating nuclear reactor based on RELAP5/SCDAPSIM/MOD3.4, Ann. Nucl. Energy 127 (2019) 215–226.

[15] D.L. Ma, et al., Supercritical water heat transfer coefficient prediction analysis based on BP neural network, Nucl. Eng. Des. 320 (2017) 400–408.

[16] D.C. Groeneveld, et al., The 2006 CHF look-up table, Nucl. Eng. Des. 237 (15–17) (2007) 1909–1922.

[17] X. Zhao, Data for: on the Prediction of Critical Heat Flux Using a Physics-Informed Machine Learning-Aided Framework, 2020. Mendeley.

[18] D. Ying, et al., Study on optimization methods of nuclear reactor radiation shielding design using genetic algorithm, Nucl. Power Eng. 37 (4) (2016) 160–164.

[19] B.S. Kim, J.H. Moon, Use of a genetic algorithm in the search for a near-optimal shielding design, Ann. Nucl. Energy 37 (2) (2010) 120–129.

[20] International Atomic Energy Agency, “Artificial Intelligence for Accelerating Nuclear Applications, Science and Technology”, https://www.iaea.org/publications/15198/artificial-intelligence-for-accelerating-nuclear-applications-science-and-technology, 2022

[21] Qingyu Huang, Shinian Peng, Jian Deng, Hui Zeng, Zhuo Zhang, Yu Liu, Peng Yuan , “A review of the application of artificial intelligence to nuclear reactors: Where we are and what’s next”, *Heliyon,* Volume 9, Issue 3, March 2023, https://www.sciencedirect.com/science/article/pii/S2405844023010903?via%3Dihub

References (cont'd)

[22] Nilsson, N. J.. The Quest for Artificial Intelligence. Cambridge University Press, 2009

[23] Ram Shankar Siva Kumar, David O'Brien, Kendra Albert, Salome Viljoen, Jeffrey Snover, "Failure Modes in Machine Learning", https://securityandtechnology.org/wp-content/uploads/2020/07/failure_modes_in_machine_learning.pdf, November, 2019

[24] Zhan, Xinhui; Sun, Heshan; and Miranda, Shaila M., "How Does AI Fail Us? A Typological Theorization of AI Failures" (2023). Forty-Fourth International Conference on Information Systems, Hyderabad, India 2023,. https://aisel.aisnet.org/icis2023/aiinbus/aiinbus/25

[25] Dario Amodei et al., «Concrete Problems in AI Safety", arXiv:1606.06565v2 , https://arxiv.org/pdf/1606.06565, 2016

[26] Rey Reza Wiyatno, Anqi Xu, Ousmane Dia, and Archy de Berker, "Adversarial Examples in Modern Machine Learning: A Review", arXiv:1911.05268v2 [cs.LG] 15 Nov 2019, https://arxiv.org/pdf/1911.05268

[27] Guang Yan, Quinghoa Ye, Jun Xia, "Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond", Information Fusion, 77(2022) 29-52

[28] "Generative Design", *Design Engineering,* April, 2025

[29] Siemens Corp., Mindsphere White paper, https://resources.sw.siemens.com/en-US/white-paper-mindsphere-whitepaper/?ste_sid=c9c9c705c6967189cea3e9c8b3faec92

[30] IBM, Maximo Application Suite, https://www.ibm.com/products/maximo?utm_content=SRCWW&p1=Search&p4=43700081186703597&p5=e&p9=58700008819665194&&msclkid=dc171ab5bf5d1f955c4b72a1e0467f51&gclid=dc171ab5bf5d1f955c4b72a1e0467f51&gclsrc=3p.ds&gad_source=7

[31] Uptake: Predictive Maintenance, https://uptake.com/

[32] Tyler Ward, Kouroush Jenab, Jorge Ortega-Moody, and Selva Staub, "A Comprehensive Review of Machine Learning Techniques for Condition-Based Maintenance", *Proc. of the 71st Reliability and Maintainability Symposium,* Destin, FL, 2025

[33] Enrico Zio, "AI-Based Prognostics and Health Management for Predictive Maintenance", *Proc. of the 71st Reliability and Maintainability Symposium,* Destin, FL, 2025

[34] P. Baraldi, F. Mangili, and E. Zio, "A prognostics approach to nuclear component degradation modeling based on Gaussian Process Regression." Progress in Nuclear Energy, 2015. 78: pp. 141-154

References (cont'd)

[35] J. B. Coble, P. Ramuhalli, L. J. Bond, W. Hines, and B. Upadhyaya, "Prognostics and health management in nuclear power plants: a review of technologies and applications". 2012, Pacific Northwest National Laboratory (PNNL), Richland, WA (US)

[36] J. Ma and J. Jiang, "Applications of fault detection and diagnosis methods in nuclear power plants: A review." Progress in nuclear energy, 2011. 53(3): pp. 255-266.

[37] H. Kim, S.-H. Lee, J.-S. Park, H. Kim, Y.-S. Chang, and G. Heo, "Reliability data update using condition monitoring and prognostics in probabilistic safety assessment." Nuclear Engineering and Technology, 2015. 47(2): pp. 204-211

[35] Paschali, Magdalini, et al. "Generalizability vs. Robustness: Adversarial Examples for Medical Imaging." arXiv preprint (2018), https://arxiv.org/pdf/1804.00504,

[36] Ebrahimi, Javid, Daniel Lowd, and Dejing Dou. "On Adversarial Examples for Character-Level Neural Machine Translation." arXiv preprint (2018), https://arxiv.org/pdf/1806.09030

[37] Carlini, Nicholas, and David Wagner. "Audio adversarial examples: Targeted attacks on speech-to-text." arXiv preprint (2018), https://arxiv.org/pdf/1801.019444

[38] Jagielski, Matthew, et al. "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning." arXiv preprint, https://arxiv.org/pdf/1801.019444

[39] Fredrikson M, Jha S, Ristenpart T. 2015. Model inversion attacks that exploit confidence information and basic countermeasures

[40] Elsayed, Gamaleldin F., Ian Goodfellow, and Jascha Sohl-Dickstein. "Adversarial Reprogramming of Neural Networks." arXiv preprint (2018), https://arxiv.org/pdf/1806.11146

[41] Sharif, Mahmood, et al. "Adversarial Generative Nets: Neural Network Attacks on State-of-the-Art Face Recognition." arXiv preprint arXiv:1801.00349 (2017).

[42] Gu, Tianyu, Brendan Dolan-Gavitt, and Siddharth Garg. "Badnets: Identifying vulnerabilities in the machine learning model supply chain." arXiv preprint https://arxiv.org/pdf/ 1708.06733 (2017)

[43] Ortega, Pedro, and Vishal Maini. "Building safe artificial intelligence: specification, robustness, and assurance." DeepMind Safety Research Blog (2018).

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, et. al., "Attention Is All You Need", 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA., https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[45] Brittney Muller, "BERT 101 State Of The Art NLP Model Explained", March 2, 2022, https://huggingface.co/blog/bert-101

[46] "T5", https://huggingface.co/docs/transformers/en/model_doc/t5

References (cont'd)

[47] Ian J. Goodfellow, Jean Pouget-Abadie∗ , Mehdi Mirza, et.al., "Generative Adversarial Nets", https://arxiv.org/pdf/1406.2661

[48] Ryan C. Barron, Maksim E. Eren, Olga M. Serafimova, Cynthia Matuszek, and Boian S. Alexandrov. "Bridging Legal Knowledge and AI: Retrieval Augmented Generation with Vector Stores, Knowledge Graphs, and Hierarchical Non-negative Matrix Factorization". *Proceedings of The 20th International Conference on Artificial Intelligence and Law (ICAIL 2025*). ACM, New York, NY, USA, 2025, https://arxiv.org/pdf/2502.20364

[49] C.A. Melton, A. Sorokine, S. Peterson, Evaluating Retrieval Augmented Generative Models for Document Queries in Transportation Safety , https://arxiv.org/abs/2504.07022

[50] Jeanie Genesis, Retrieval-Augmented Text Generation: Methods, Challenges, and Applications , https://www.preprints.org/manuscript/202504.0443/v1, April, 2025

[51] Michael Karanicolas, Artificial Intelligence and Regulatory Enforcement (report to the Admin. Conf. of the U.S.), Dec. 9, 2024 https://www.acus.gov/sites/default/files/documents/AI-Reg-Enforcement-Final-Report-2024.12.09.pdf,

[52] Prantikos, K., Chatzidakis, S., Tsoukalas, L.H. *et al.* Physics-informed neural network with transfer learning (TL-PINN) based on domain similarity measure for prediction of nuclear reactor transients. *Sci Rep* **13**, 16840 (2023). https://doi.org/10.1038/s41598-023-43325-1

[53] Towards a Standard for Identifying and Managing Bias in Artificial Intelligence, NIST Special Publication NIST SP 1270, March, 2022, https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf

[54] Secure Software Development Practices for Generative AI and Dual-Use Foundation Models, NIST Special Publication NIST SP 218-A, https://airc.nist.gov/technical-reports/#nist-sp-800-218a-secure-software-development-practices-for-generative-ai-and-dual-use-foundation-models

[55] Reducing Risks Posed by Synthetic Content, NIST AI 100-4, https://airc.nist.gov/docs/NIST.AI.100-4.SyntheticContent.ipd.pdf

[56] Four Principles of Explainable Artificial Intelligence, NISTIR 8312, https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8312.pdf

Backup: Summary Description of Artificial Intelligence Technologies

Artificial Intelligence Technologies Descriptions

The term “Artificial Intelligence” refers to algorithms and statistical models that enable computers to perform tasks without being explicitly programmed. AI systems learn and improve from experience by identifying patterns and making decisions based on data. In contrast, conventional programs where computers perform tasks by processing explicit instructions.

The diagram on the previous chart shows a taxonomy of artificial intelligence (AI) algorithms. The top level of the taxonomy distinguishes machine learning from evolutionary computing. Machine Learning (ML) uses learning or training, where training is the search for weights of learning function parameters to minimize an objective function (e.g., the difference between the actual value of a data point and its predicted value). Evolutionary computing uses algorithms inspired by natural evolution. They are used to solve optimization and search problems. However, evolutionary computing models have not been used in nuclear applications of AI as far as we could determine in our literature search and will not be discussed further in the following pages

We have defined the following ML categories:

- **1.1 Supervised Learning**: The model is trained on labeled data, i.e., each training data sample includes both the data value and the known output. Common tasks include classification (predicting a category such as a failure mode) and regression (predicting a continuous value such as a power level).
- **1.2 Unsupervised Learning**: The model is trained on unlabeled data and tries to find hidden patterns or intrinsic structures in the input data. Common tasks include clustering and association.
- **1.3 Reinforcement Learning**: The model learns by interacting with an environment and receiving feedback in the form of rewards or penalties. The goal is to learn a strategy that maximizes cumulative rewards over time.
- **1.4 Generative AI**: Models that can generate new content based on an input data set. Generative models span supervised and unsupervised learning, and are therefore placed in their own category

The following pages discuss some of the more important techniques in each of these categories. However, the discussion is not exhaustive, and new developments in ML (and in AI more generally) are occurring rapidly – so rapidly that one of the most important for a for publications of new methods is Arxiv, which hosts early (often non-peer reviewed but nevertheless widely cited articles).

Where not specifically referenced, the text in the descriptions of the AI methods was initially generated by the OpenAI GPT-4o Large Language Model (LLM) and subsequently manually edited and verified.

1.1 Supervised Machine Learning Techniques (1/5)

1.1.1 Neural networks:

A neural network – or more specifically, an artificial neural network (ANN) -- is a software simulation of the way biological neural networks in the brain process information. The network is composed of interconnected layers of nodes, or "neurons," which work together to recognize patterns, make decisions, and solve complex problems. ANNs consist of 3 types of layers:

- **Input Layer**: This layer receives the initial data or inputs. Each neuron in this layer represents a feature or attribute of the input data.
- **Hidden Layers**: One or more such layers sit between the input and output layers. They perform various transformations and computations called “activation functions” on the inputs received from the previous layer. In the simplest networks, there is only one hidden layers, but the depth (number of hidden layers) can be much higher. The depth impact the network's ability to capture intricate patterns.
- **Output Layer**: This layer produces the final output of the network. The number of neurons in this layer corresponds to the number of possible outcomes or predictions.

Each connection between neurons has an associated weight, which determines the strength and direction of the signal. During training, the network adjusts these weights based on the prediction errors using techniques like backpropagation and gradient descent.

There are multiple types of neural networks. Some of the common categories are:

- **Feedforward Neural Networks (FNN)**: This is simplest type of artificial neural network. Information moves in one direction—from input nodes, through hidden nodes (if any), to output nodes. FNNs are used for basic pattern recognition and classification tasks.
- **Convolutional Neural Networks (CNN)**: CNNs are specialized for processing data with grid-like topology, such as images. They use convolutional layers that apply filters to detect features like edges, textures, and objects. CNNs are widely used in image and video recognition, medical image analysis, and more.
- **Recurrent Neural Networks (RNN)**: RNNs handle sequential data by having connections that loop back. They can maintain a form of memory, making them suitable for tasks like language modeling, time-series prediction, and speech recognition. Variants include Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), which help mitigate issues like vanishing and exploding gradients.
- **Autoencoders**: Autoencoders are neural networks for unsupervised learning used for tasks like dimensionality reduction, feature learning, and data compression. They consist of an encoder that compresses the input into a latent-space representation and a decoder that reconstructs the input from this representation. Autoencoders are further described in the section on Generative AI
- **Generative Adversarial Network (GAN)**: GANs consist of two networks, a generator and a discriminator, that are trained simultaneously. The generator creates fake data, while the discriminator evaluates its authenticity. They are useful for generating realistic images, videos, and other types of data. GANs are further described in the section on Generative AI
- **Radial Basis Function Networks (RBFNs)**: RBFNs use radial basis functions (RBFs) as activation functions (an RBF value is determined by the distance from a central point, known as a "center; the most common RBF is a Gaussian density function with a zero standard deviation). RBFNs are typically used for function approximation, time-series prediction, and control.

1.1 Supervised Machine Learning Techniques (2/5)

1.1.2 Deep Neural Networks

As noted earlier, neural networks consist of input, hidden, and output layers. Deep neural networks have many (two to twenty) hidden layers. The benefit is allowing them to learn hierarchical representations of data.

The key characteristics of deep learning using neural networks include:

- **Multiple Hidden Layers**: Unlike traditional neural networks, deep learning models have several hidden layers, which enable them to capture intricate patterns and representations in data.
- **Feature Learning**: Deep learning models can automatically learn and extract features from raw data, reducing the need for manual feature engineering.
- **Scalability**: Deep neural networks can handle large-scale datasets and complex tasks, making them suitable for applications such as image and speech recognition, natural language processing, and more

Specialized architectures, such as Convolutional Neural Networks (CNNs) for image processing and Recurrent Neural Networks (RNNs) for sequential data, have been developed in deep learning to address specific types of problems more effectively.

Deep learning has become highly popular due to its success in many fields, driven by advancements in computational power, availability of large datasets, and improvements in training algorithms.

Supervised Machine Learning Techniques (3/5)

1.1.3 Support Vector Machines

A Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification (finite discrete outputs). However, it can also be used for and regression (continuous value outputs) tasks. The key idea behind SVM is to find the optimal hyperplane that best separates the data into different classes. An SVM is particularly useful in situations where the number of dimensions exceeds the number of samples. Here are some of the key concepts:

- **Hyperplane**: a boundary that separates different classes in the feature space. For a two-dimensional space, this would be a line; for higher dimensions, it becomes a plane or hyperplane.
- **Support Vectors**: These are the data points that are closest to the hyperplane and are critical in defining its position and orientation. The optimal hyperplane is the one that maximizes the margin between the support vectors of different classes.
- **Margin**: The margin is the distance between the hyperplane and the nearest data points from each class (support vectors). SVM aims to maximize this margin, thereby achieving better generalization on unseen data.
- **Kernel Functions**: When data is not linearly separable in its original feature space, SVM can apply a kernel function to map the data into a higher-dimensional space where it becomes linearly separable. Common kernel functions include:
    1. **Linear Kernel**: Used when the data is linearly separable.
    2. **Polynomial Kernel**: Maps the data into a higher-dimensional polynomial space.
    3. **Radial Basis Function (RBF) Kernel**: An RBF is a function whose value depends on the distance from a center point, line, or plane. An RBF kernel maps the data into an infinite-dimensional space, often used for non-linear classification.
- **Regularization**: SVM includes a regularization parameter (C) that controls the trade-off between maximizing the margin and minimizing classification error. A higher value of C aims to classify all training examples correctly, while a lower value allows some misclassification to achieve a larger margin.
- **Applications**: SVM is widely used in various domains such as text classification, image recognition, bioinformatics, and more, due to its versatility and effectiveness in high-dimensional spaces.

1.1 Supervised Machine Learning Techniques (4/5)

1.1.5 Decision Trees

Decision trees are a type of machine learning algorithm used for both classification and regression tasks. They are models that make decisions based on a series of (yes/no) questions (tests or decisions), with the result being a tree structure. The tree consists of nodes, a decision or test whose outcome is determined by the value of a certain feature, branches, representing the outcome of the decision or test, and leaf nodes represents a category or a regression value.

There are two types of decision trees: regression trees for predicting continuous numerical values and classification trees, for the predicting categories.

The following are the key aspects:

- **Tree Structure**: The regression tree is structured as a binary tree, where each internal node represents a decision based on the value of a certain feature, and each leaf node represents a predicted continuous value (for regression trees) or category (for classification trees).
- **Splitting Criteria**: The tree is built by splitting the data at each node based on criteria that minimize a cost function such as the mean squared error (MSE) or mean absolute error (MAE). The goal is to create splits (data subsets) that result in the most homogeneous subgroups in terms of the target variable.
- **Prediction**: To make a prediction for a new data point, the regression tree starts at the root and traverses down the tree based on the feature values of the data point until it reaches a leaf node. The value at the leaf node is the predicted continuous value for that data point.
- **Pruning**: To avoid overfitting, regression trees can be pruned. Pruning involves removing branches that have little importance and do not contribute significantly to the model's accuracy, thereby simplifying the model.

Regression trees are useful for their interpretability and ability to handle non-linear relationships between features and the target variable.

1.1.6 Random Forest

A random forest combines the predictions of multiple decision trees to improve predictive performance and control overfitting. It is an ensemble learning method and can be used for classification, regression, and other tasks. The underlying concept of random forests is that a group of weak learners (individual trees) can be combined to create a strong learner (the random forest). Each tree in the ensemble (“forest”) is trained on a different random subset of the training data. The subsets are samples with replacement (known as bootstrap aggregating, or bagging). This means that each tree is trained on a different portion of the data, promoting diversity among the trees. To further increase diversity, a random subset of features is used for splitting at each split, rather than using all features. This randomness helps create trees that are less correlated with each other, enhancing the robustness and generalization ability of the model. For classification tasks, the forest aggregates the outputs of individual trees through majority voting. For regression tasks, it averages the predictions of the individual trees. The aggregation or averaging helps reduce variance and improve accuracy.

1.1 Supervised Machine Learning Techniques (5/5)

Random Forests (continued)

Random forests are less prone to overfitting than individual decision trees, especially when there are many trees. They are also capable of handling large datasets with many features (higher dimensionality). By considering feature subsets at each split random forests can provide estimates of feature importance, helping in understanding which features contribute the most to the prediction. They are particularly useful where the relationship between the features and the target variable is complex and non-linear

Random forests are widely used due to their flexibility, robustness, and ability to handle both classification and regression problems effectively.

1.1.7 Gradient Boost Machines

A Gradient Boosting Machine (GBM) is an ensemble machine learning technique used for regression and classification tasks. It builds the model in a stage-wise fashion by sequentially adding weak learners, typically decision trees, and optimizing them to correct the errors made by the previous models. Here are the key characteristics of GBM:

- **Boosting Technique**: Unlike bagging (used in random forests), boosting focuses on combining weak learners sequentially, where each new model attempts to correct the errors of the previous models.
- **Gradient Descent**: The "gradient" in gradient boosting refers to the use of gradient descent to minimize the loss function. At each stage, the algorithm fits a new model to the residual errors (the difference between the actual and predicted values) of the combined ensemble of previous models.
- **Additive Model**: The overall model is built by adding the outputs of the weak learners. Mathematically, the next model in the sequence is chosen to minimize the prediction error, often by following the negative gradient of the loss function.
- **Loss Function**: GBM can be tailored for various types of problems through the choice of the loss function. Common loss functions include mean squared error for regression and logistic loss for classification.
- **Learning Rate**: A key parameter in GBM is the learning rate, which scales the contribution of each new model. A lower learning rate often requires more trees to be added but can lead to better generalization.
- **Regularization**: GBM includes regularization techniques to prevent overfitting, such as limiting the maximum depth of the trees, specifying minimum samples per leaf, and using subsampling methods.

Popular implementations of gradient boosting include XGBoost, LightGBM, and CatBoost, each offering enhancements and optimizations to the basic gradient boosting framework. In addition to PHM, GBM is used finance, healthcare, and marketing, due to its high predictive performance and flexibility.

1.2 Unsupervised Machine Learning Techniques (1/3)

1.2.1 Clustering

Clustering is used to group similar data points together based on certain characteristics or features, without using predefined labels. The goal is to identify inherent structures or patterns within the data. Clustering helps to identify natural groupings or clusters within the data. This can be particularly useful in exploratory data analysis to understand the distribution and relationships in the dataset. For PHM, its application would be Identifying unusual patterns or outliers in data, which can be used for fault detection.

Common Clustering Algorithms include:

- **K-Means**: Divides the data into a predetermined number of clusters (k) by minimizing the variance within each cluster. It iteratively assigns data points to the nearest cluster centroid and updates the centroids.
- **Hierarchical Clustering**: Builds a hierarchy of clusters either through an agglomerative approach (bottom-up) where each data point starts as its own cluster and merges with others, or a divisive approach (top-down) where the entire dataset starts as one cluster and splits into smaller clusters.
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**: Forms clusters based on the density of data points in the feature space, allowing it to identify clusters of varying shapes and sizes and handle noise.
- **Gaussian Mixture Models (GMM)**: Assumes that the data is generated from a mixture of several Gaussian distributions and uses probabilistic models to find the clusters.

.Because clustering is unsupervised and does not have predefined labels, evaluating the quality of the clusters can be challenging. Common evaluation metrics include:

- **Silhouette Score**: Measures how similar a data point is to its own cluster compared to other clusters.
- **Davies-Bouldin Index**: Evaluates the average similarity ratio of each cluster with the one that is most similar to it.
- **Elbow Method**: Used with K-Means to determine the optimal number of clusters by plotting the explained variance as a function of the number of clusters and looking for an "elbow" point.

1.2 Unsupervised Machine Learning Techniques (2/3)

1.2.2 Association Rule Learning

Association rule learning is a machine learning method used to discover interesting relationships, patterns, or associations within large sets of data. It is primarily utilized in the context of market basket analysis (collections of items that appear together frequently in transactions) and then generating association rules from these item sets. The strength of these rules is typically measured using metrics such as:

- **Support**: The proportion of transactions in the dataset that contain the itemset.
- **Confidence**: The likelihood that the presence of one item leads to the presence of another item.
- **Lift**: The ratio of the observed support to that expected if the items were independent.

For example, in a retail context, an association rule might reveal that customers who buy bread are also likely to buy butter. This kind of insight can be used for diagnostics and prognostics

1.2.3 Dimensionality reduction

Dimensionality reduction reduces the number of random variables under consideration. This simplification is achieved by obtaining a set of principal variables or features that preserve essential information while eliminating redundancy and noise. The primary goals of dimensionality reduction are to improve computational efficiency, enhance model performance, and make data visualization more manageable. Dimensionality reduction can be used in can be used in supervised and unsupervised learning

There are two main types of dimensionality reduction techniques:

1. **Feature Selection**: This approach involves selecting a subset of the original variables based on certain criteria, such as variance, correlation, or relevance to the target variable. Common methods include:
    1. Filter methods (e.g., Chi-square test, Information Gain)
    2. Wrapper methods (e.g., Recursive Feature Elimination)
    3. Embedded methods (e.g., LASSO, Ridge Regression)
2. **Feature Extraction**: This involves transforming the data from a high-dimensional space to a lower-dimensional space, often by creating new features that are combinations of the original ones. Popular techniques include:
    1. Principal Component Analysis (PCA)
    2. Linear Discriminant Analysis (LDA)
    3. t-Distributed Stochastic Neighbor Embedding (t-SNE)
    4. Autoencoders (a type of neural network)

By reducing dimensionality, these techniques help address issues related to overfitting, increased complexity, and poor model performance.

1.2 Unsupervised Machine Learning Techniques (3/3)

1.2.4 Autoencoder Models

An autoencoder is a type of artificial neural network used for unsupervised learning. Its primary purpose is dimensionality reduction, i.e., elimination of unimportant features of the data set (see Dimensionality Reduction heading under Unsupervised Machine Learning Techniques). An autoencoder consists of an encoder and a decoder. The encoder compresses the input data (with the full set of original features) into latent variables in a latent space (reduced feature) which represents the input data with fewer features, while the decoder reconstructs the input data from the compressed (reduced feature) latent variables.

Training the autoencoder network involves minimizing the “loss function”, i.e., the difference between the input data and its reconstruction from the decoder. The resultant optimization identifies the important features of the input data (“feature learning”). Autoencoders are widely used in tasks such as image denoising, anomaly detection, and data compression. They can be used for unsupervised learning labeled data is scarce, as they do not require labels to learn the underlying structure of the data.

1.3 Reinforcement Learning

In reinforcement learning, a software agent (learner or decision maker) learns to make decisions by performing actions in an environment (the external system with which the agent interacts) with the goal of maximizing cumulative reward from the feedback on the outcome. Unlike supervised learning, where the model is trained with input-output pairs, reinforcement learning relies on a feedback loop where the agent receives rewards or penalties based on the actions it takes.

Key components of reinforcement learning include:

- **State**: A representation of the current situation of the agent within the environment.
- **Action**: The set of all possible moves the agent can make.
- **Reward**: The feedback received from the environment after performing an action, which can be positive or negative.
- **Policy**: A strategy used by the agent to determine the next action based on the current state.
- **Value Function**: A function that estimates the expected cumulative reward for a given state or state-action pair.

The agent's goal is to learn a policy that maximizes the total reward over time. This involves exploring different actions to discover their effects and exploiting known actions that yield high rewards. Popular algorithms in reinforcement learning include Q-learning, Deep Q-Networks (DQN), and Proximal Policy Optimization (PPO).

Reinforcement learning has been successfully applied in robotics,, autonomous driving, and resource management.

1.4 Generative AI (1/4)

Generative AI models can generate new content from existing data. This content could include text, images, music, and even video. Unlike traditional AI, which often focuses on recognizing patterns and making decisions based on existing data, generative AI creates new data that is similar to the input data it was trained on. Categories of generative AI models include

- Generative adversarial networks (GANs) are widely used for creating realistic images and videos. These models have a wide range of applications, from creative arts and entertainment to more practical uses like data augmentation and automated content creation.
- Variational Autoencoders use probability distributions rather than the fixed values of autoencoders to produce new content.
- Large Language Models (LLMs) are a subset of generative AI models focused on natural language. A widely known example of generative AI is OpenAI's GPT-4, a language model that can produce human-like text based on the prompts it is given.

These categories are discussed in greater detail in the following charts.

1.4.1 Generative Adversarial Networks

Generative Adversarial Networks (GANs) [47, 40] are a class of machine learning frameworks that generate new data with the same characteristics as a given dataset. GANs consist of two neural networks:

- Generator: This network creates data samples, attempting to mimic the real data distribution. It starts with random noise and transforms it into data that resembles the training set.
- Discriminator: This network evaluates the data samples, distinguishing between real data from the training set and fake data produced by the generator. It provides feedback to the generator to improve the quality of the generated samples.

The generator and discriminator are trained simultaneously through adversarial processes. The training process involves the generator trying to produce data that can fool the discriminator, while the discriminator learns to better identify fake data. Over time, the generator improves its ability to create realistic data samples, and the discriminator becomes more adept at distinguishing between real and fake data.

GANs are used in applications such as image synthesis, video generation, and creating art. They are known for their ability to produce high-quality, realistic outputs, but they can be challenging to train due to issues like mode collapse and instability.

1.4 Generative AI (2/4)

1.4.2 Variational Autoencoder Models

Variational Autoencoder models (VAE) ARE variant of the standard autoencoder (discussed in the section on unsupervised learning) that uses probabilistic mappings from the input data to the latent space rather than fixed values (see autoencoder description for a definition of input data, latent variables, and latent spaces). Unlike traditional autoencoders, VAEs are designed to generate new data samples that are similar to the input data.

Like Autoencoder models, VAEs include an Encoder, Latent Space, and a Decoder. As noted in the previous paragraph, the Autoencoder maps the input data to a single point or vector in the latent space. In a VAE, the encoder outputs parameters of a probability distribution (usually a Gaussian distribution) that represents the latent variables. The VAE uses a specific loss function that combines a reconstruction loss (measuring how well the decoder reconstructs the input data) with a regularization term (usually the Kullback-Leibler divergence) that ensures the learned latent space distribution is close to a prior distribution (often a standard normal distribution).

VAEs are widely used in applications that require generating new data samples, such as image synthesis, anomaly detection, and data imputation. They are particularly valued for their ability to interpolate between data points in the latent space, producing smooth transitions and variations of the input data.

1.4 Generative AI (3/4)

1.4.3 Large Language Models

1.4.3.1 Transformers

Transformer based models ("Transformers") are a type of neural network architecture introduced in the paper "Attention is All You Need" by Vaswani et al. [44]. They are primarily used for natural language processing tasks but has also been adapted for other domains. The transformer architecture is based on the mechanism of self-attention, which allows the model to weigh the significance of different words in a sentence when making predictions. This self-attention mechanism enables transformers to capture long-range dependencies and relationships in data more effectively than previous models like recurrent neural networks (RNNs).

Transformers consist of an encoder and a decoder (similar to autoencoders). The encoder processes the input data and creates a representation, while the decoder uses this representation to generate the output. One of the key innovations of transformers is their ability to process input data in parallel, which significantly speeds up training and inference compared to sequential models like RNNs.

Transformers are versatile and can be used in both supervised and unsupervised learning, They have become the foundation for many state-of-the-art models in Natural Language Processing, including Bidirectional Encoder Representations from Transformers (BERT) [45], Generative Pretrained Transformers (GPT, see below), and T5 [46], among others. These models have been applied to natural language processing tasks, such as translation, summarization, and question answering.

1.4.3.2 Generative Pre-trained Transformers

A Generative Pre-trained Transformer (GPT) is a type of artificial intelligence model designed for natural language processing tasks. It is based on the transformer architecture (see previous chart). The GPT model is characterized by its ability to generate human-like text by predicting the next word in a sequence, given a prompt or context. The "pre-trained" aspect refers to the model being initially trained on a large corpus of text data to understand language patterns, grammar, and context. This pre-training phase allows the model to learn a general understanding of language, which can then be fine-tuned for specific tasks such as translation, summarization, or question-answering.

1.4 Generative AI (4/4)

1.4.3.2 Generative Pre-trained Transformers (continued)

GPT models have a Transformer Architecture that utilizes self-attention mechanisms to weigh the importance of different words in a sentence, allowing for better context understanding and parallel processing. The have a Generative Capability, i.e., they can generate coherent and contextually relevant text, making it useful for applications like chatbots, and content creation. They can be scaled up with more parameters and training data to improve performance, as seen in the OpenAI iterations of Chat GPT moving from GPT-2, GPT-3, GPT-4, and beyond.

GPT models are applicable to a wide range of language tasks without task-specific architectures, relying on fine-tuning for specific applications. This versatility has significantly advanced the field of natural language processing by providing tools for understanding and generating human language.

1.4.3.3 Retrieval Augmented Generation

Retrieval Augmented Generation (RAG) is a technique that combines retrieval-based and generative models. In more concrete terms, an RAG system combines a library of externally stored documents with the internal data of a large language model. The key advantage of RAG is that it allows the model to access a vast amount of external knowledge in addition to the LLM internal model. This additional access enhances the its ability to provide information that can be referenced by the user. This approach is particularly useful in scenarios where the model needs to answer questions or provide insights that require specific and detailed knowledge beyond its pre-existing training data and increases the confidence that the data does not contain erroneous inferences (“hallucinations”).

The RAG involves a three-step process: retrieval, augmentation, and generation. In retrieval, a search of relevant data or documents in an external document repository is initiated by a user query or prompt. This could be a question, a topic for content generation, or a specific information request. The LLM searches through the external data repository looking for regulations, documents, articles, reports, or other relevant sources. The system then filters the search results to prioritize the most relevant and highest quality information. In augmentation, the identified documents are retrieved and the system integrates the external data with its internal training data. The model analyzes the context of the retrieved data to determine how it can enhance the generative process. In generation, the final response is created. Using both the retrieved data and its internal training data, the model generates a response. The LLM assesses whether the content meets standards of accuracy and relevance, and then produces the final response to the user.