NERC
NORTH AMERICAN ELECTRIC
RELIABILITY CORPORATION

Artificial Intelligence and Machine Learning in Real-Time System Operations

White Paper – Revision 1

November 2024

RELIABILITY | RESILIENCE | SECURITY

3353 Peachtree Road NE
Suite 600, North Tower
Atlanta, GA 30326
404-446-2560 | www.nerc.com

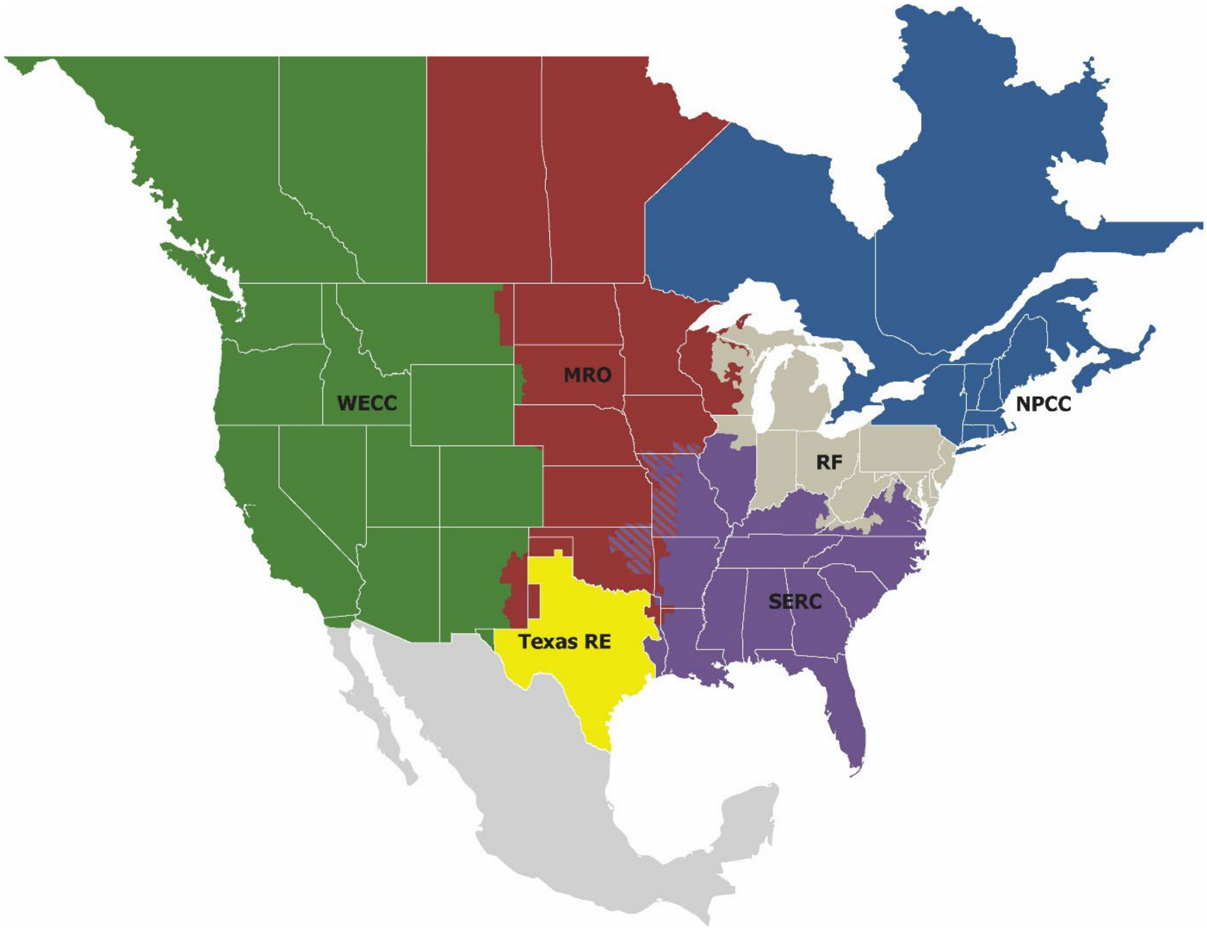
Table of Contents

Preface

Electricity is a key component of the fabric of modern society and the Electric Reliability Organization (ERO) Enterprise serves to strengthen that fabric. The vision for the ERO Enterprise, which is comprised of NERC and the six Regional Entities, is a highly reliable, resilient, and secure North American bulk power system (BPS). Our mission is to assure the effective and efficient reduction of risks to the reliability and security of the grid.

Reliability | Resilience | Security
Because nearly 400 million citizens in North America are counting on us

The North American BPS is made up of six Regional Entities as shown on the map and in the corresponding table below. The multicolored area denotes overlap as some load-serving entities participate in one Regional Entity while associated Transmission Owners/Operators participate in another.



MRO	Midwest Reliability Organization
NPCC	Northeast Power Coordinating Council
RF	Reliability First Corporation
SERC	SERC Reliability Corporation
Texas RE	Texas Reliability Entity
WECC	WECC

Executive Summary

While Artificial Intelligence (AI), Machine Learning (ML), and Data Science have been studied and developed for decades, technological growth and public attention have recently ballooned. This has led to a tremendous amount of research and new solutions in the marketplace, affecting nearly every aspect of work and personal life. Innovations and discourse continue to evolve rapidly as excitement grows and investment and research branch into new areas.

Within the realm of real-time electric power operations, there is also recognition of the increasing complexity and complicatedness of the BPS given ongoing changes, with several new use cases that stretch the assumptions of the system (e.g., increasing concerns around cyber aspects, excess solar flowing onto the transmission system, significant load growth for electric vehicle charging, growing power requirements for AI/ML, cryptocurrency mining on blockchains, and other data center operations). The BPS is the backbone of North America's energy infrastructure. It is crucial for security and economic stability across both the continent and the nation and underpins our daily lives.

Managing the real-time reliability of the system requires control room operators to possess increasing levels of cognition, attention, vigilance, knowledge, and abstract reasoning, invariably leading many to consider new AI/ML solutions. Because the BPS is the planet's most complex sociotechnical system (a system involving complex humans and complex systems with complex interactions between them), many considerations are needed to minimize the risks to the system.

This document is intended for decisionmakers, regulators, and end users of these technologies, particularly in real-time operations. It is not helpful to assert whether these technologies should be used in real-time operations, as surveys and interviews of key stakeholders throughout the industry show that this is already happening. The "genie" cannot be put back in the bottle (and this document does not assert that it should). Instead, this document provides guidance on the kinds of questions that one should ask about these technologies to thoroughly understand what they are capable of and what kinds of changes are needed to implement them properly.

Previous technologies to come to market have fallen into typical patterns, leading to initial "bumpy" implementations with unexpected risks or adverse events. This document offers a path for real-time operations—in which such adverse events are intolerable—in an attempt to ensure that AI/ML technologies can be implemented in a way that maximizes the chances of a successful, reliability-enhancing deployment.

There is strong recognition that many organizations across the industry are already considering AI/ML applications and making a variety of decisions from actively trying to avoid them (e.g., blocking Generative Pre-trained Transformer (GPT) access from work computers and enacting policies about information security) to embracing them (e.g., leveraging better customer call tracking, ensuring strengthened asset health, and predicting real-time operational parameters, such as wind generation, solar generation, and load).

The surface area of AI/ML technologies, both present and future, is vast. This document focuses on currently available technologies that are built, trained, and deployed to deal with specific situations and would not work outside the domain of their training (e.g., a solar generation predictor could not be relied upon to predict wind generation), generally called narrow AI (or sometimes, weak AI). This includes recent areas of rapid growth, including the ability to generate new content (with Generative AI algorithms such as GPT).

Several aspects of the intersection of electric power operations and AI/ML are out of scope of this white paper, which specifically focuses on the implementation and usage of AI/ML technologies in real-time operations by system operators. Out-of-scope items include the following:

- AI/ML technologies as a rapidly growing load on the BPS
- The intersection of AI/ML and cloud technologies
- AI/ML use in near-real-time applications, such as planning
- Aspects of supply chain, such as generative AI usage by vendors in systems design

This document is divided into several sections:

- An overview of the technologies currently within the AI/ML space and distinctions between the types
- An overview of the human factors associated with real-time control room operators working alongside AI/ML and the kinds of preparations that are needed to increase the likelihood of desirable outcomes
- An overview of an anonymized survey and interviews from decisionmakers within the industry to provide a snapshot of the kinds of use cases and considerations that these entities are investigating
- An overview of AI/ML technologies and the roles that they can play in real-time operations
- An overview of cybersecurity risks that AI/ML technologies pose to the reliable real-time operations of the BPS

The implementation of AI/ML systems is one of the many changes and challenges faced by society. Of particular interest in this context is the fact that humans are changing at the same time in response to shifts in our environment. For example, as our performance on the abstract reasoning components of IQ tests continues to rise over time (Flynn, 2013), our ability to collaborate with technology in more abstract and advanced ways allows growing human and technological strengths to tackle more advanced system challenges in real-time operations. Nevertheless, there is also a history of new, advanced technologies initially failing, not because of technological failures specifically but because of an insufficient focus on the interactions between humans and the system. This paper also advocates that these new systems can avoid those early pitfalls if they are properly designed, implemented, and used.

Implementing AI/ML systems into real-time operations requires a strong relationship with the humans to allow the humans to effectively maintain, operate, and question the accuracy of the systems' responses, both in real-time operations and in facilitating the identification of new ways of developing, testing, and deploying the systems to reduce the risks of novel errors. These systems should not be implemented under a mindset that sees "humans as hazards" and seeks to avoid or disintermediate human involvement but rather one that seeks to bring the strengths of humans and systems together to achieve even higher levels of effectiveness and reliability. Similarly, successful implementation also requires recognition that, as support/aid tools, AI/ML systems will provide operators more mental bandwidth to monitor and strengthen the reliability of the system—but not if that bandwidth is removed through the addition of more switch-tasking and distractions in the system operators' daily work.

This white paper is a product of collaboration between NERC, ResilientGrid, the NERC Energy Management System Working Group (EMSWG), and the Real-Time Operating Subcommittee (RTOS). The paper focuses on the benefits, risks, and opportunities of AI/ML in real-time system operations and provides a high-level overview of AI/ML in those operations. The technology known as AI/ML is not new and dates back to the 1950s. However, the recent growth and ubiquity of AI/ML is impacting an increasing range of sectors and is expected to affect global productivity. Assessing AI/ML's impacts on the energy sector, both as an enabler and as a new source of risks, requires a thorough understanding of the relevant technology.

The intention of this white paper is to ensure that system operators are involved in the decision-making process when AI/ML is being tested, implemented, and used. The operator should have the final input on the decision for AI/ML-generated actions that are to be taken. One of the goals of this paper is to seek opportunities to help industry make informed decisions, enhance reliability, and respond swiftly to changing grid conditions.

Ultimately, this white paper asserts that AI/ML systems—if properly scoped, developed, implemented, and monitored and enacted with proper training and continuous improvement—can augment the efforts of dedicated, engaged, and talented real-time system operators to increase the reliability, robustness, and resilience of the BPS.

Chapter 1: Key Terms

Within this document, several terms describe aspects of the BPS, humans, and AI and ML systems. Terminologies may be somewhat different within diverse disciplines, so a brief lexicon is offered here.

General Terms

- **Human-AI Teaming:** In some disciplines, other phrases like "Human-Computer Interaction," "Human-Robot Interaction," "Artificial Co-Pilot," or "Human-Centered Cyber-Physical System" are also used to describe a human subject matter expert and some kind of advanced technological system working together. Within the context of this document, "system" refers to an AI/ML system.

- **Narrow AI**: This refers to the AI systems that are well proven at a particular task but cannot be generalized beyond that. For example, a system that is able to predict consumer behavior as it relates to energy consumption on-peak is not likely to provide reasonable results when estimating other activities (e.g., EV charging behavior or water use).

- **Generative AI**: This refers to advanced AI/ML systems that can generate new content. These capabilities currently extend to the generation of text, imagery, audio, video, and computer code and use technologies such as GPT.

- **Mental Model:** Mental models incorporate our assumptions and generalizations and other ways that we understand the world and how it works. We use mental models to interpret the situations in which we find ourselves, identify what we see as our goals, and influence or make decisions. Mental models can also be thought of in the same way that models in AI/ML systems (e.g., in a pre-trained transformer) are built, updated, and accessed for pattern recognition, prediction, and decision making.

- **Human and Organizational Performance (HOP):** HOP (formerly Human Performance Improvement; HPI) is the science and application of humans working individually, in groups, and in/throughout organizations to maximize the benefits of having humans doing the work (e.g., abstract reasoning, collaboration, creativity) while minimizing the risks of humans doing work (e.g., human errors, cognitive biases). This encompasses fields of science like Human Factors Engineering, Cognitive Systems Engineering, Resilience Engineering, and Naturalistic Decision Making. People in these fields seek to identify and maximize human potential and reliability as part of advanced, reliable, and resilient complex systems.

Complex vs. Complicated

- **Complex (e.g., Snowden & Boone, 2007):** A complex system has many relationships between its pieces that are hidden or difficult to view. For example, a specific change to one component may create a completely unexpected change elsewhere in the system that would have been difficult or impossible to determine prior to the change. Within the realm of critical infrastructures, these unexpected/emergent changes often occur during high-stress, high-stakes situations. Complex systems are also likely to continually change, further making predictions difficult.

- **Complicated (e.g., Snowden & Boone, 2007):** This is a system that has many relationships that can be completely separated from each other, allowing for problem-solving in one area that enables others to be ignored. For example, on a vehicle, it is safe to assume that the suspension of the vehicle as it drives on a dirt road cannot be altered by changing the vehicle's windshield wiper fluid levels. The human mind and engineering disciplines are particularly well-suited to solving these kinds of problems, supporting the reduction of scope/degrees of freedom and thereby allowing for easier and more linear solving of problems. However, approaching complex problems as though they are complicated often leads to unexpected (negative) outcomes.

Reliability vs. Resilience vs. Robustness

Many definitions with subtle differences for reliability, robustness, and resilience exist throughout the industry (e.g., Zissis, 2019, Jones, 2021). For the purposes of this white paper, the definitions of Hollnagel, 2009 are used, as they are the seminal definitions, aligning with Resilience Engineering principles and associated human factors design principles.

- **Reliability:** This is a current measure of the system's state, which is measured by how well it meets its operational requirements (for example, those specified in NERC, 2013). For example, a transmission system that is operating in its expected configuration, allowing for successful generation onto and load off of the system, can be considered reliable. This reliability is degraded by customers being off-line and unable to receive power.

- **Robustness:** This is a measure of the degree of perturbation necessary to bring the system from a reliable state to a non-reliable state, which includes its ability to recover from foreseen and mitigated circumstances. For example, the more that a system has equipment failing or off-line (losses of redundancy/security) or is operating at or beyond its operational limits (losses of adaptive capacity), the less robust it is likely to be. For example, a power system that would lead to widespread relay trips following a single fault has poor robustness.

- **Resilience:** This is a measure of the degree to which the system and the humans that support it are able to recover and/or adapt the system from a non-reliable state back into a reliable state, often in response to unforeseen circumstances. For example, Black Start restoration is a strong example of resilience, and the more effectively and quickly a utility can restore the system, the more resilient the system is. Resilience can also be thought of as the inverse of brittleness.

- **Antifragility:** This is a special case of resilience that occurs when humans (potentially aided by technology) perform "what-if" analysis, identifying threats to reliability and ways of minimizing those risks, or dealing with them better should they occur.

Information vs. Data

- **Data:** Data is the fundamental element that human senses perceive, and computers receive. Effectively, data is something (e.g., a measurement) without context. For example, the number 100 can be entered into a computer or read by a person, but its meaning is unclear without additional data, processing, or comprehension (e.g., good for a test score percentage, bad for a person's body temperature in Fahrenheit; a boiling point for water in Celsius).

- **Information**: Information is a more complex concept that includes context, allowing one to infer what something means (e.g., good or bad), predict what will occur next, or run scenarios against to determine the best course of action. The human brain tends to "chunk" information together to form higher levels of information (e.g., an area code may be encoded as the city it is from rather than three distinct numbers), or an expert chess player may be able to encode the configuration of most of the chess piece layout into a single "chunk" (and an associated sense of risks and responses). Because humans can hold only 7 ± 2 of these "chunks" of information (Miller, 1956), the more context held within one chunk, the more valuable it is. The more expert a person is at something, the better they are able to encode information into fewer chunks.

Additional HOP Concepts

- **Complex Sociotechnical System (Hollnagel, Woods & Leveson, 2006):** A system in which complex humans, systems, and societies interact with one another. Due to their complexity, these systems cannot be reliably decomposed and will always have hidden interdependencies that are not consistently visible.
- **Cognitive Systems Engineering (Hollnagel & Woods, 1983):** The design of human-centered complex systems that considers both the strengths and limitations of humans in cognitive work, ensuring that humans are in control of the system and able to understand its state.
- **Human Factors Engineering:** The science of engineering systems that includes cognitive work and physical work to maximize human function and limit risks associated with human limitations.
- **Resilience Engineering (Hollnagel, 2009)**: The design of systems to have the ability to adjust functioning prior to, during, and following disturbances, whether expected or unexpected, to maintain a reliable state.
- **Adaptive Capacity (Hollnagel, Woods & Leveson, 2006):** The capability of a system (human, technological, or both) to respond/adapt in order to maintain or return to reliable operations. These capacities are often taxed by factors such as stress, switch-tasking, and data overload in humans and limited data, processing power, or control capabilities on the technical side.
- **Naturalistic Decision Making (Klein, 2008):** The recognition that work, as it is actually performed in real-world scenarios, cannot be accurately understood or predicted using lab-based or simplified applied research. Therefore, in order to understand and strengthen complex operations, working with actual experts, whether in interviews, observations, or high-fidelity simulations, provides the most accurate and likely-to-be-successful answers.

Chapter 2: Technology Overview

AI and ML have become integral to modern technology, transforming industries and everyday life. Additionally, Data Science is crucial in extracting insights from data, driving decision-making, and supporting the development of AI and ML models. Deep Learning (DL), a subset of ML, has further revolutionized the field with its ability to handle vast amounts of data and complex patterns. Figure 2.1 illustrates the relationship between AI, ML, DL, and Data Science.

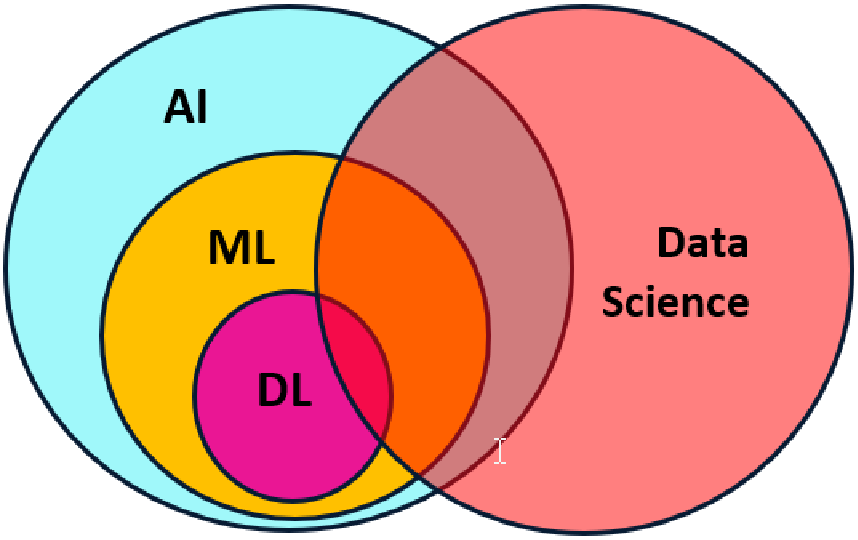


Figure 2.1: AI, ML, DL, and Data Science

AI alignment is a critical aspect of AI development, ensuring that AI systems operate safely, ethically, and in a manner consistent with the values of society and the application for which the AI is used. This chapter provides a general technology overview of AI, ML, DL, Data Science, and AI alignment, detailing fundamental concepts and methods.

Artificial Intelligence

AI refers to the simulation of human intelligence in machines, allowing them to perform tasks that typically require human intelligence, such as problem-solving, learning, reasoning, perception, and language understanding. Developers of AI systems aim to create systems that can exhibit "intelligent" behavior and adapt to different situations, making the systems capable of accomplishing tasks without explicit human programming. AI can be categorized into two main types (Figure 2.2): Generative AI and Predictive AI.

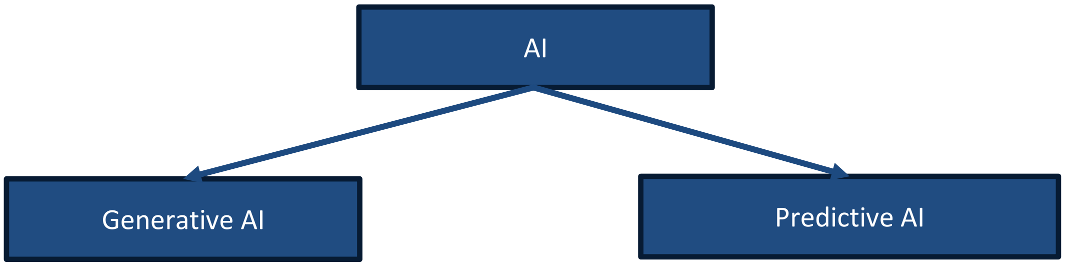


Figure 2.2: Types of AI

Generative AI

Generative AI focuses on creating new content based on existing data. This form of AI can produce text, images, music, and other media, mimicking the style and coherence of human-generated content.

Key Technologies and Methods

Key technologies and methods in Generative AI include the following:

- **Generative Adversarial Networks (GAN)**

  GANs consist of two neural networks—a generator and a discriminator—that contest with each other. The generator creates data samples, while the discriminator evaluates their authenticity. Through this adversarial process, the generator improves its ability to produce realistic data. GANs are used in image synthesis, creating realistic human faces, enhancing video game graphics, and even drug discovery.

- **Variational Autoencoders (VAE)**

  An autoencoder is a type of neural network used to learn efficient coding of unlabeled data (unsupervised learning; Kramer, 1991). An autoencoder learns two functions: an encoding function that transforms the input data and a decoding function that recreates the input data from the encoded representation. VAEs are a type of autoencoder designed to generate new data points similar to the training data. They work by encoding input data into a latent space and then decoding it back to reconstruct it, allowing for the generation of new but similar data. VAEs are used in generating new text, creating variations of images, and augmenting data for training other AI models.

- **Transformer Models**

  Transformer architectures, such as GPT-3 and GPT-4, leverage attention mechanisms to handle long-range dependencies in data, particularly in natural language processing (NLP). These models can generate coherent and contextually relevant text. Transformer models are used in chatbots, automated content creation, translation services, and summarization tools.

Advantages of Generative AI

Using this technology offers several benefits. The advantages of Generative AI may include the following:

- **Creativity and Innovation**

  Generative AI can create new, original content, which can be used in fields such as art, music, and design. This can lead to new forms of creativity and innovation, especially when paired with human experts.

- **Personalization**

  Generative AI can create personalized content based on individual preferences and behaviors. This can enhance user experiences in areas like online shopping, entertainment, and education.

- **Augmentation**

  Generative AI can generate synthetic data, which can be used to augment real data in scenarios where data collection is challenging, or privacy concerns exist.

Disadvantages of Generative AI

Since Generative AI is trained on a set of data and can only generate content based on the information that it is fed, the use of poor data or data containing unlicensed content can lead to copyright infringement, privacy breaches, bias, and non-compliance. Generative AI can be used to create deepfakes or synthetic media, which can be used for misinformation or fraud, raising ethical and legal issues. Other challenges, such as hallucinations, are discussed further in this document. Organizations using Generative AI should establish AI governance standards (e.g., NIST, 2023) to mitigate these risks.

Predictive AI

Predictive AI focuses on forecasting future outcomes based on historical data. This type of AI is crucial for making data-driven decisions across various domains.

Key Technologies and Methods

Key approaches in Predictive AI include the following:

- **Regression Analysis**

  Regression models predict continuous outcomes by learning the relationship between input variables and the output variable. Linear regression is a simple yet powerful technique, while more complex methods like polynomial regression and support vector regression (SVR) handle non-linear relationships. Regression analysis is used in financial forecasting, demand prediction, and risk assessment.

- **Classification**

  Classification algorithms predict categorical outcomes by learning from labeled data. Techniques such as logistic regression, decision trees, and neural networks are commonly employed for classification tasks. Classification is essential in spam detection, medical diagnosis, and image recognition.

- **Time Series Analysis**

  Time series models, including Autoregressive Integrated Moving Average (ARIMA) and Long Short-Term Memory (LSTM) networks, analyze sequential data to predict future values based on past trends. Time series analysis is pivotal in stock market prediction, weather forecasting, and resource allocation.

Advantages of Predictive AI

Like Generative AI, Predictive AI offers unique benefits. These advantages include the following:

- **Improved Decision-Making**: Provides insights into better strategic decisions
- **Efficiency:** Optimizes processes and resource allocation
- **Competitive Advantage:** Identifies opportunities and threats in advance

Disadvantages of Predictive AI

Since Predictive AI is trained on large amounts of data to forecast, more and/or higher-accuracy data is needed to ensure its efficacy. No future events can be predicted with absolute certainty, and aspects of context may not necessarily be reliably attended to. Any organization using this technology must recognize that, as in all cases, technology has limitations.

Machine Learning

ML is a subset of AI that focuses on enabling machines to learn from data and improve their performance over time without explicit programming. In other words, instead of instructing a machine to perform a specific task, it is provided with data to learn patterns and relationships within the data to give predictions, make decisions, or identify patterns. ML algorithms are broadly categorized into three types: Supervised Learning, Unsupervised Learning, and Reinforcement Learning, as depicted in **Figure 2.3**.

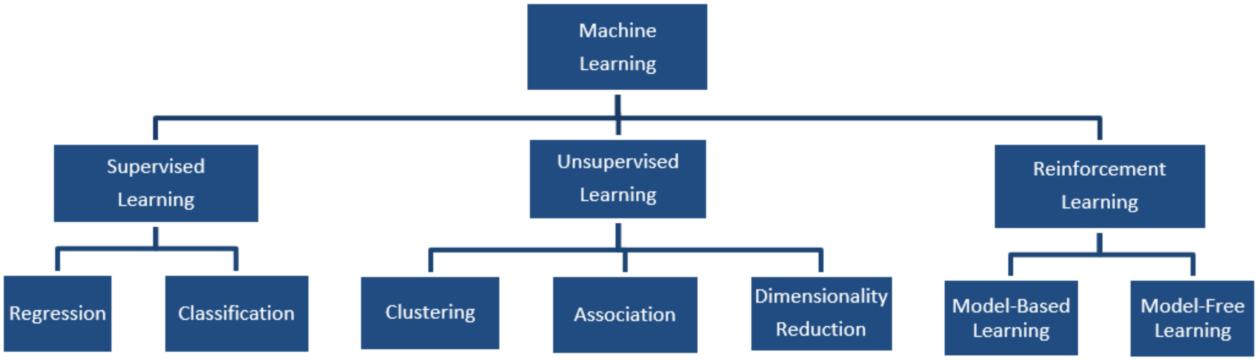


Figure 2.3: A Simple Illustration of ML Algorithms

In supervised learning, a dataset includes its desired outputs (or *labels*) so that a function can calculate an error for a given prediction. The supervision comes when a prediction is made and an error (actual vs. desired) is produced to alter the function and learn the mapping.

In unsupervised learning, a dataset does not include a desired output, meaning that there is no way to supervise the function. Instead, the function attempts to segment the dataset into "classes" so that each class contains a portion of the dataset with common features.

In reinforcement learning, the algorithm attempts to learn actions for a given set of states that lead to a goal state. An error is provided not after each example (as is the case for supervised learning) but instead on receipt of a reinforcement signal (such as reaching the goal state). This behavior is similar to human learning, where feedback is not necessarily provided for all actions but only when a reward is warranted.

Figure 2.4 illustrates learning models for ML algorithms.

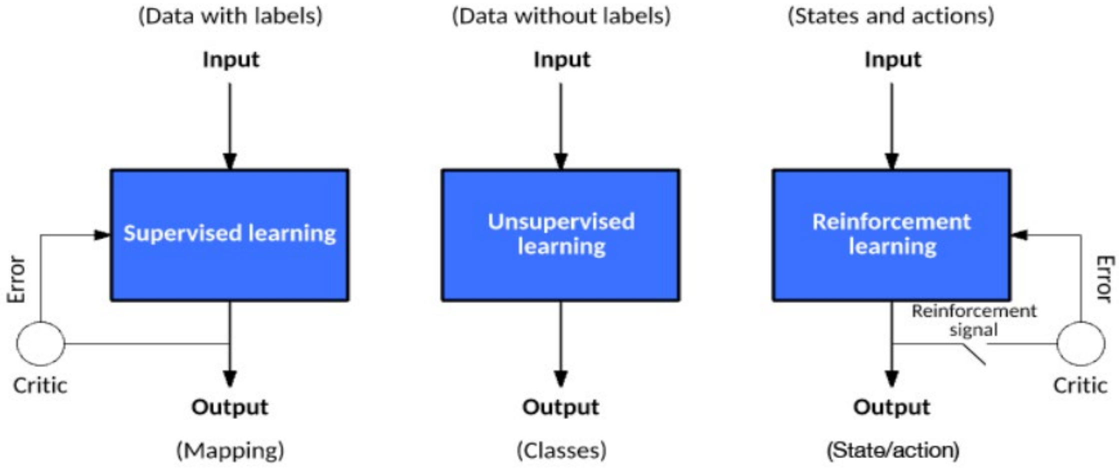


Figure 2.4: Learning Models for ML Algorithms (Jones, 2017)

Supervised Learning

Supervised Learning (when an algorithm is trained on a labeled dataset and provided with the input data and corresponding correct outputs) is a highly accurate and efficient way of making predictions. The model learns from the labeled examples and generalizes to make predictions on new, unseen data. Supervised learning is classified into two categories of algorithms: regression and classification.

- **Regression**
  Regression is a type of supervised learning that is used to predict continuous values. Regression algorithms learn a function that maps from the input features to the output value. Common regression algorithms include the following:
  - **Linear Regression**
    Linear regression models the relationship between a dependent variable and one or more independent variables using a linear equation. It is simple and interpretable, making it a foundational technique in supervised learning.
  - **Decision Trees**
    Decision trees split the data into subsets based on the value of input features, creating a tree-like model of decisions. They are intuitive and useful for both classification and regression tasks.
- **Classification**
  Classification is a type of supervised learning that is used to predict categorical values. Classification algorithms learn a function that maps from the input features to a probability distribution over the output classes. Common classification algorithms include the following:
  - **Neural Networks**
    Neural networks consist of layers of interconnected nodes (neurons) that process input data. DL, a subset of neural networks with many layers, has achieved remarkable success in tasks such as image and speech recognition.
  - **Support Vector Machines (SVM)**
    SVMs find the optimal hyperplane that separates different classes in the feature space. They are effective in high-dimensional spaces and for classification tasks with clear margin separation.

Unsupervised Learning

Unsupervised learning is used for exploratory tasks to discover hidden patterns or relationships in the data with no goal or target in mind. It is when the algorithm is given an unlabeled dataset and tries to find patterns, structures, or relationships in the data without explicit guidance. Generally, there are three types of unsupervised learning: clustering, dimensionality reduction, and association.

- **Clustering**
  Clustering is a technique for exploring raw, unlabeled data and breaking it down into groups (or clusters) based on similarities or differences. It is used in a variety of applications, including customer segmentation, fraud detection, and image analysis. Clustering algorithms split data into natural groups by finding similar structures or patterns in uncategorized data. Common algorithms include K-means, hierarchical clustering, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN).
- **Dimensionality Reduction**
  Dimensionality reduction is an unsupervised learning technique that reduces a dataset's number of features or dimensions. More data is generally better for ML, but it can also make it more challenging to visualize the data. Dimensionality reduction extracts important features from the dataset, reducing the number of irrelevant or random features present. This method uses principal component analysis (PCA) and singular value decomposition (SVD) algorithms to reduce the number of data inputs without compromising the integrity of the properties in the original data.

- **Association**
  Association rule learning is a rule-based approach used to reveal interesting relationships between data points in large datasets. Unsupervised learning algorithms search for frequent if-then associations—also called rules—to discover correlations and co-occurrences within the data and the different connections between data objects. The *a priori* algorithms are the most widely used for association rule learning to identify related collections of items or sets of items. However, other types, such as Eclat and FP-growth algorithms, are used.

Reinforcement Learning

Reinforcement learning is learning to make decisions to achieve a specific goal. It is based on an agent interacting with an environment to act and receive feedback through rewards and penalties. The agent's goal is to learn the best actions to take in different states of the environment to maximize its cumulative reward over time. Reinforcement learning is beneficial in scenarios with unknown optimal actions, and the agent needs to learn by trial and error. Generally, there are two types of reinforcement learning: model-based learning and model-free learning.

- **Model-Based Learning**
  Model-based learning is typically used when environments are well-defined and unchanging and where real-world environment testing is difficult. The agent first builds an internal representation (model) of the environment. Once the model is complete, the agent simulates action sequences based on the probability of optimal cumulative rewards. It then further assigns values to the action sequences themselves. The agent thus develops different strategies within the environment to achieve the desired end goal.

- **Model-Free Learning**
  Model-free learning is used when the environment is large, complex, and not easily describable. It is also ideal when the environment is unknown and changing and environment-based testing does not have significant downsides. The agent does not build an internal model of the environment and its dynamics. Instead, it uses a trial-and-error approach within the environment. It scores and notes state-action pairs—and sequences of state-action pairs—to develop a policy.

  Q-learning is a model-free reinforcement learning algorithm that learns the value of actions in each state. The Q-value represents the expected future rewards of taking a specific action in a given state, guiding the agent's policy. Q-learning is used in game playing, robotics, and autonomous vehicles.

Deep Learning

DL is a subset of ML that uses multi-layered neural networks, called deep neural networks, to simulate the complex decision-making power of the human brain (Holdsworth & Scapicchio, 2024). Deep neural networks consist of multiple layers of neurons, with each neuron receiving inputs, applying a weight, and passing the output through an activation function. The network learns by adjusting these weights based on the error in predictions. These networks can automatically discover intricate patterns and relationships within the data, making them robust for tasks like image and speech recognition, NLP, and decision-making.

To train a neural network, there are two propagations, as depicted in Figure 2.5**.**

- **Forward Propagation**
  Forward propagation is where input data is fed through a network in a forward direction to generate an output. The data is accepted by hidden layers and processed, as per the activation function, and moves to the successive layer. The forward flow of data is designed to avoid data moving in a circular motion, which does not generate an output.

  During forward propagation, pre-activation and activation take place at each hidden and output layer node of a neural network. The pre-activation function is the calculation of the weighted sum. The activation function is applied, based on the weighted sum, to make the neural network flow non-linearly using bias.

- **Back Propagation**
  In the back propagation, the flow is reversed—starting by propagating the error to the output layer until reaching the input layer passing through the hidden layer(s). The back propagation algorithm is the set of steps used to update network weights to reduce network errors.

Together, forward propagation and back propagation allow a neural network to make predictions and correct any errors accordingly. Over time, the algorithm becomes gradually more accurate.

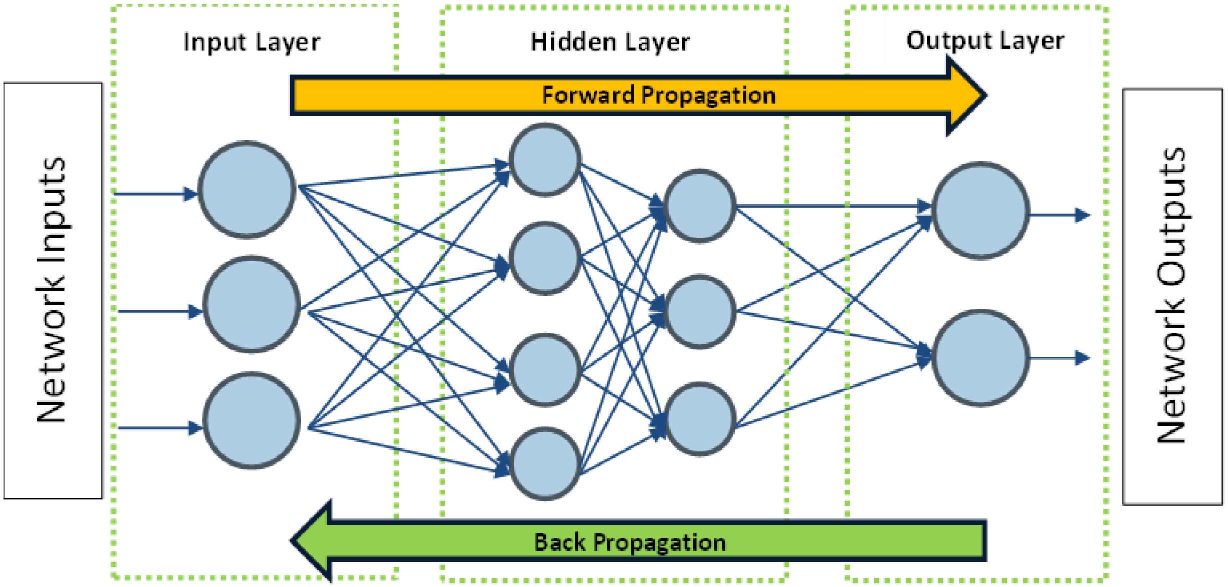


Figure 2.5: A Simple Illustration of DL Architecture

Critical concepts in DL include the following:

- **Neural Networks**
  Neural networks consist of layers of neurons, with each neuron receiving inputs, applying a weight, and passing the output through an activation function. The network learns by adjusting these weights based on the error in predictions.
- **Convolutional Neural Networks (CNN)**
  CNNs are specialized neural networks designed for processing structured grid data like images. They use convolutional layers to learn spatial hierarchies of features automatically and adaptively.
- **Recurrent Neural Networks (RNN)**
  RNNs are designed for sequence data and can use their internal state (memory) to process sequences of inputs. LSTM networks and Gated Recurrent Units (GRU) are popular RNN variants that address the vanishing gradient problem.
- **Transformers**
  Transformer models use self-attention mechanisms to process sequential data, allowing them to handle long-range dependencies effectively. These models have achieved state-of-the-art results in NLP tasks.

Data Science

Data Science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data. It encompasses a wide range of techniques from statistics, computer science, and domain-specific knowledge. The main components of Data Science include data collection, data validation, data cleaning, data analysis, and data visualization and reporting, illustrated in **Figure 2.6**.

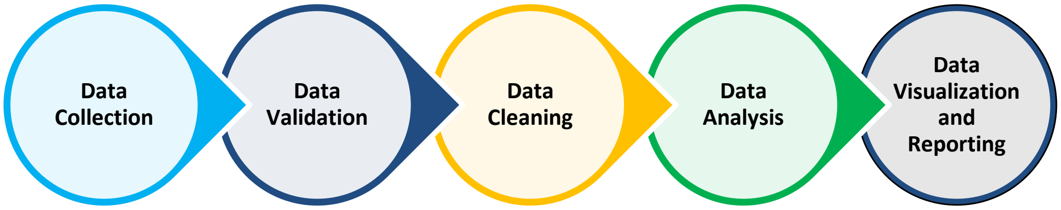


Figure 2.6: Main Components of Data Science

Data Collection

Data collection involves gathering raw data from various sources, including databases, web scraping, sensors, and surveys. Data Collection is the actual process of gathering data on targeted variables identified as data requirements. The emphasis is on ensuring correct and accurate data collection, which means that the correct procedure was taken, and appropriate measurements were adopted and that the maximum efforts were spent to ensure the data quality. Key steps include the following:

- **Identifying Data Sources**

  Sources can be internal (e.g., company databases) or external (e.g., public datasets, Application Programming Interfaces (API).

- **Data Ingestion**

  Data ingestion is the process of importing data into a storage system for analysis. This can involve batch processing or real-time streaming.

Data Validation

Data validation refers to the process of ensuring the accuracy and quality of data. It is implemented by building several checks into a system or report to ensure the logical consistency of input and stored data. Generally, data validation can include the following:

- Data type validation
- Range and constraint validation
- Code and cross-reference validation
- Structured validation
- Consistency validation
- Relevancy validation

Data Cleaning

Data cleaning involves removing errors and inconsistencies from data to improve its quality. This step is crucial for ensuring accurate and reliable analysis. Key tasks include the following:

- **Handling Missing Data**
  Techniques include removing missing values, imputing missing values with statistical measures, or using advanced methods like k-nearest neighbors (KNN) imputation.
- **Removing Duplicates and Errors**
  This involves identifying and removing duplicate records and correcting data entry errors.
- **Normalization and Standardization**
  This involves normalizing data to a common scale or standardizing it to have a mean of zero and a standard deviation of one.

Data Analysis

Data analysis involves applying statistical and computational techniques to extract meaningful insights from the data. This step often employs ML algorithms and statistical models. Key methods include the following:

- **Exploratory Data Analysis (EDA)**
  EDA uses summary statistics and visualizations to understand the data's structure, identify patterns, and detect anomalies.
- **Statistical Analysis**
  This involves applying statistical tests and models, like t-tests, chi-square tests, and regression analysis, to infer relationships and test hypotheses.
- **Machine Learning**
  This involves implementing ML algorithms to predict outcomes and uncover hidden patterns.

Data Visualization and Reporting

Data visualization involves representing data in graphical or pictorial form to make insights more accessible and understandable. Effective visualization can highlight trends, patterns, and outliers, aiding decision-making. Key aspects include:

- **Types of Visualizations**
  - **Bar and Line Charts**: Used for comparing categories or showing trends over time
  - **Histograms**: Used for showing the distribution of a single variable
  - **Scatter Plots**: Used for identifying relationships between two variables
  - **Heatmaps**: Used for representing data values in a matrix with varying colors
- **Visualization Tools**
  - **Tableau**: A powerful tool for creating interactive and shareable dashboards
  - **Power BI**: A Microsoft tool for business analytics and data visualization
  - **Matplotlib, Seaborn, Plotly**: Python libraries for creating static, animated, and interactive visualizations
  - **D3.js**: A JavaScript library for producing dynamic, interactive data visualizations in web browsers

Data reporting systematically documents and presents data, information, or findings in a structured, textual, or tabular format. Typically used in business, research, or academic contexts, reporting communicates essential details, analyses, and outcomes to a specific audience. Reports provide a narrative context to the data, explaining its significance, implications, and recommendations. They aim to offer a comprehensive and detailed account of a subject, making them essential tools for decision-making, performance evaluation, and information sharing within organizations and across various fields.

The choice between data visualization and reporting is a crucial decision in data communication. Data visualization conveys insights quickly and effectively, making it ideal for identifying patterns and trends. Conversely, reporting offers a detailed narrative context, making it suitable for comprehensive information sharing. The decision should hinge on the nature of the data, the preferences of the audience, and the communication objectives. Combining these two approaches can provide a balanced strategy for enhancing data-driven decision-making, offering the best of both worlds in conveying information effectively.

AI Alignment

AI alignment refers to the effort to ensure that AI systems act in accordance with human values and goals. This is crucial because AI could potentially cause harm by diverging from human interests as it becomes more advanced and autonomous. Ensuring alignment helps mitigate these risks and ensures that AI systems contribute positively to society.

There are several approaches to AI alignment. One key strategy involves designing AI systems with robust value alignment mechanisms, ensuring that they prioritize human values and goals. Additionally, ongoing research into alignment methods, transparency, and ethical considerations can help guide the development of safe and beneficial AI. Collaboration among researchers, policymakers, and industry stakeholders is crucial for effectively addressing alignment challenges.

The challenges in AI alignment include the following:

- **Value Specification**: Defining human values in a way that can be effectively encoded into AI systems is complex and subjective.
- **Scalability**: Ensuring alignment as AI systems become more powerful and autonomous poses scalability challenges, especially considering the potential for unintended consequences.
- **Alignment Verification**: It is difficult to verify whether AI systems are truly aligned with human values, especially as they become more complex and opaque.
- **Adversarial Dynamics**: Adversarial actors may attempt to manipulate AI systems for their own ends, leading to misalignment or harmful outcomes.
- **Robustness**: AI systems need to be designed to withstand, as much as possible, unforeseen circumstances and edge cases to maintain alignment in diverse environments.

Addressing these challenges requires interdisciplinary collaboration, ongoing research, and careful consideration of ethical implications throughout the development and deployment of AI technologies. These factors, as well as others related to the quality of inputs to AI/ML systems (e.g., data lineage, provenance, and quality), are covered in the next chapter.

Chapter 3: Human Factors in AI/ML

As BPS real-time operations continue to grow ever more complex and complicated, the role of human operators becomes even more critical. Introducing AI/ML technologies, especially predictive and generative technologies, can drastically change how operators build and maintain situational awareness, make decisions, and ensure the reliability of the grid. Quality implementation, testing, and training should allow humans working with AI/ML systems to significantly improve the reliability and resilience of the BPS.

Throughout history, humans have employed new technologies to enhance their work, but the early implementations of these technologies have often proven bumpy. For example, the introduction of autopilot in aviation led to a significant increase in fatal airline crashes due to design choices that unintentionally pushed humans "out of the loop," decreasing their situational awareness and adaptive capacity to respond in emergencies (Smith, et al., 1997, Endsley & Kris, 1995, Merlo, 2012).

Because major failures are virtually unacceptable in real-time electric power operations, a focus on the human factor considerations of AI/ML technologies is important to ensure that these technologies, when used in real-time operations, increase the reliability and resilience of the BPS as intended.

Throughout this section, several aspects of human-machine teaming are discussed, with focuses on the following:

- How AI/ML systems can be designed to better interface and work with humans
- How aspects of human performance (e.g., attention, cognitive bias) may affect human interactions with AI/ML systems
- How integrating these systems will benefit from more advanced/newer approaches to training and simulations

Some of these aspects, such as human error and cognitive bias, are discussed with the recognition that identification and mitigation (e.g., with HOP aspects as part of training) of these risks will increase the reliability of the BPS and the humans working in real-time operations. Any system, whether technological or human, can have risks and failures, which can only start to be mitigated once recognized. Overall, organizations with more advanced human performance maturity will be able to adapt, implement, and manage these systems more effectively and with lower risk.

Human-Machine Teaming

Joint Cognitive System

Because critical real-time operations work cannot be conducted by humans or technology alone, a different lens is required to recognize the system that emerges from the two groups working together. This approach, the joint cognitive system, transcends views such as "humans working with machines" into "humans and machines working together" in the highly complex environment where this work occurs.

Joint activity has at least four basic requirements that apply to both humans and technology equally (Hollnagel & Woods, 2005):

- Entering into an agreement, called a basic compact, that the participants intend to work together
- Having a reasonable degree of predictability in behavior
- Having the ability to be directed, to receive and appropriately respond to instructions
- Maintaining a "common ground" (common beliefs, understandings, and communications)

Contract Between Humans and AI/ML

One of the most critical factors in Human-Machine Teaming is the contract between the two parties. In many places, the lack of clear understanding of the contract, whether on the part of the developers/testers or users, can lead to undesirable outcomes. Users should clearly understand:

- What the interface is between the human and the machine and how it is implemented
- How the human should expect to have information conveyed to them on an ongoing basis
    - Including confidence intervals for predictions, identifications, anomaly detections, etc.
- How the human should expect to be alerted with time-sensitive information
- How the human should report incorrect predictions and conclusions and how the machine should receive them and respond
- How the machine should interface with the human for needed clarification, coordination, and validation
    - Including the expectations of the human in response (e.g., how they should respond and how quickly)

Trust and Confidence in AI/ML

Due to the critical nature of real-time operations, a system operator's trust in the system, the quality of its data, and the quality of AI/ML performance is important in determining how the operator will interact with the system. For example, a system operator who distrusts data flowing from a Remote Terminal Unit (RTU) will significantly change the operator's speed, steps, and decisions on the Supervisory Control and Data Acquisition (SCADA) system. In the same way, an operator who is given recommendations from an AI/ML system, or who reviews generated text that appears to be incorrect, will not only change behavior in that moment but also likely in the near future until that confidence is restored.

As an example, users' ability to see a visualization of how input data is transformed in an AI/ML system improved their perception of the system's reliability (Kerr, et al., 2007). However, automation bias (Mosier & Skitka, 1999) makes people more likely to assume the correctness of a machine's output rather than a human's. Overcoming these and several other biases and patterns to target a consistent "trust but verify" approach to reviewing and taking action on AI/ML data is necessary for a successful outcome. The more the systems incorporate diverse data, provide transparency and explainability, and solicit feedback for continuous improvement, the more the users will be able to approach reasonable levels of trust.

Prediction and Understanding

While a great deal of research has been done in attempts to create technological systems that work like neurons in the human brain, the core processes in computer systems tend to work quite differently than human brains. For example, the creation and behavior of an AI "neural network" is substantially different from a section of the human neocortex in terms of its layout, connectivity, complexity, dynamic nature, and propagation of information throughout the whole brain. Similarly, even though the relationships that many of these models can predict may seem, at least on the surface, to resemble natural active conversations with an expert human possessing significant understanding, these systems do not process—nor responded to—data in the same ways as humans.

Within the realm of real-time operations, these technologies are being considered and implemented in environments with little margin for error, in an infrastructure that is identified as the most critical in America (White House, 2013). Because we rely on humans to ultimately understand decisions and their implications, despite these technologies' advancements, they still cannot be tested by an operator to determine correct reasoning or understanding in the way one operator can ask questions of another. Consider Searle's thought experiment, the Chinese Room (Searle, 1980 & Cole, 2020), in which a set of rules can be followed to produce a correct answer without the agent performing the work understanding the inputs, outputs, or implications. This is why hybrid teamwork between humans and AI can be so powerful if done correctly, bringing the unique strengths of each group to the benefit of the BPS.

It is often possible to detect accurate, statistically significant relationships between things that are, in fact, not causative or predictive but rather related to other “black box” factors. For example, there is a statistically significant relationship between the amount of ice cream sold in a city and the probability of a murder that day, but human understanding recognizes this is not a complete picture (and certainly not something that is controllable by, for example, banning ice cream).

Beyond the question of whether an AI/ML tool truly understands something, humans tend to interpret the results of these systems in ways that may not be accurate. For example, humans are not only likely to assume understanding, but they are also likely to assume that the path that an AI/ML tool took in arriving at an answer is the same that they would have taken even though this is not always the case. Therefore, a system that addresses the “explainable AI” problem (e.g., as described in Minh, at al., 2022) can help humans avoid these biases and strengthen the monitoring of, and partnership with, these AI/ML systems.

These systems are typically not designed/trained with some of the advanced processes that humans would use when analyzing a situation (e.g., a system operator looking at a post-contingency violation will understand beyond a generic violation to what is specific to this one, such as whether a hospital or government building would be affected). As scenarios move from the commonly occurring (e.g., 50th percentile on a Gaussian distribution) to complex rare situations (e.g., things that happen 0.1% of the time), these differences between AI/ML systems and human experts become more obvious, and critical.

Many of the cognitive biases covered in this document identify ways in which humans may incorrectly perceive AI/ML systems as matching or exceeding human intelligence and understanding, leading to an over-reliance on these systems, over-trust in these systems at the expense of human experts, and a general decline in attention and situational awareness on the part of human users. While utilities can employ many solutions, key solutions of note are covered in the following sections:

- Training for AI/ML System Use
- Training for Wide-Area Knowledge and Expertise
- Human/Organizational Performance
- Training with High-Fidelity Simulations
- Technological Level – Explainability
- Addressing Bias in Training Models
- Recognition of Acceptable Input Ranges
- Lack of Generalizability/Agility to New Data or Scenarios
- Anticipation of System Failures/“Trust but Verify”

Modes of Operation

When evaluating AI/ML systems for use in real-time operations, understanding the types of behaviors in which the systems will engage, and the degree of autonomy that they have to do this work, is of critical importance. Systems may be designed with one intention in mind, but human behavior will lead to a different result.

For example, some automotive “autopilot” systems are designed with the expectation that the driver will be fully attentive and engaged in the operation of the vehicle, ready to override the system as needed. Nevertheless, these systems often design the driver “out of the loop,” subtly messaging the driver that their attention is not necessary. Over time, this can lead to full driver disengagement, which at times of critical sensor failures or misinterpretation of data (e.g., a flashing school speed limit sign) can lead to an accident. As such, identifying the system’s intended mode, and its activities to actively enforce that level of interaction between human and system, helps humans to properly assess the risk profile of the tool.

These modes of operation also include explicit decisions (e.g., the specifications and interactions expected by designers or engineers implementing the system) and implicit ones (e.g., the relative ease by which an operator may start quickly pressing an "OK" button, leading to a habit that, over long periods of time, especially in high-stress, high-stakes situations, may lead to a different mode of operation). This is why "human-in-the-loop" operation requires explicit design decisions to ensure that it is not in name only.

Active Control

Active control systems are ones in which the AI/ML system is empowered to make decisions, actuate devices, or engage in other actions, either with or without final approval of a human. An example of this type of system is an airplane's autopilot, which, once activated, is able to make decisions to maintain an airplane in flight within normal flight parameters.

These systems are particularly susceptible to "out of the loop syndrome," in which humans lose situational awareness because they receive little information about the state of the system and current operational parameters. This is why early autopilot systems led to an increase in fatal airline crashes. Most active control systems are intentionally designed with a range of operational parameters, and, when the systems recognize that those parameters are exceeded, they cease operations and cede control to a human. A human is likely to be handed control in some of the most dynamic, difficult, and/or risky situations, meaning that a loss of situational awareness is especially dangerous here (Merlo, 2012).

Co-Pilot

Co-pilot systems are those in which the AI/ML system plays more of a partnership role with the human. Because both parties are important in achieving success, the effectiveness of the system is determined by the effectiveness of the technology, the humans, and the interactions between the two.

Decision-Making Support

These systems are ones in which the AI/ML systems play a role that is more likely to be directed by a human, or to provide human(s) with information that is needed to aid humans in making decisions or in acquiring/maintaining/strengthening situational awareness. Examples of these systems are those that provide intelligent alarm management or that are able to predict/forecast the outcomes following a decision or alert a human that a complex data pattern has been found or that an expected one was not.

Technological Level

Explainability

Among the key factors in human-machine teaming is in an understanding of why a conclusion is reached, a pattern is recognized, or a course is recommended. Humans in particular are likely to assume that machines (and other humans as well) are arriving at conclusions using the same information and processes that they themselves follow, which are not always correct. This occurs especially when a high level of confidence and/or trust is placed on the system, likely because of high levels of performance in the past. These assumptions can become particularly problematic because AI systems are strong at detecting patterns but not to the same degree of understanding as humans. This problem can take many forms, such as incorrect assumptions of causation, a lack of comprehension of how components are interrelated, or a failure to consider factors that were not explicitly provided in training data (e.g., regulatory or political concerns).

Given the criticality of real-time operations, an expert real-time operator/support engineer's ability to investigate further to ensure that the right answers are being given is critical. This is similar to other controls already employed in Energy Management Systems (EMS) (e.g., operators and engineers can review State Estimator (SE) residuals and receive alarms when SE does not converge or has residuals above specified thresholds).

Take for example a generative AI tool that drafts text for reports based on system conditions. If trained on previous reports, it could assume that a current event is due to the same cause as previous events (e.g., a unit that has been derated due to fuel supply issues several times in the past month may be derated now due to another reason). If not trained properly, AI may be unable to distinguish between initial reports and updates that include root cause and not recognize the implications of stating root cause in an initial report.

Organizational Level

Vendor Transparency

While AI/ML technologies have been in use for decades, recent advancements in technological capabilities and new use cases for the technology have made it inevitable that, like with all emerging technologies, lessons learned and opportunities for continuous improvement of the technology will arise along with the recognition that bugs, challenges, or biases may affect the functioning of the technology. On the vendor side, it is therefore critical that vendors share information about these issues publicly and transparently, ideally in forums that are shared across their customers/industries, at the very least for their active users. These vendors must ensure that they have aligned their incentives internally (e.g., how they manage and reward identifications and reporting of issues) and externally (how they respond to dissemination of that information outside the company).

This is similar to other efforts around incentivizing public disclosure, such as NERC lessons learned or in near-miss database systems. Some of these systems require aspects like well-aligned key performance indicators (KPI), the anonymity of reporter and vendor, and a third party (e.g., NASA for the Aviation Safety Reporting System) to ensure that this information continues to flow. It is already the case that several AI/ML products for a particular function have several offerings from different vendors, so the identification of a risk or challenge in one may help other vendors identify and remediate similar risks in their own products. As such, these efforts will increase the reliability of these tools across vendors.

Integrity in Marketing

Due to the rapid growth of AI/ML capabilities in recent years, some of the participants interviewed see the technology as infallible or a direct replacement for human comprehension and decision making. Because of this, companies marketing their AI/ML platforms should be particularly careful in how they communicate, internally and externally, the capabilities of their products. Some high-level guidelines are recommended, as follows:

- Ensuring that the contract/mode of interaction between humans and the system is clearly stated (e.g., whether it is designed to replace human comprehension and decision making, like a fully autonomous vehicle, or work side-by-side with a human, like a decision-making support tool)
- Ensuring that the tool's capabilities and limitations are clearly defined and specified. For example, based on how the system is trained, what types of problems can it solve and what cannot reliably be solved?

Because AI/ML platforms are a major area of discussion and source of excitement and hype, vendors must behave ethically and responsibly in communicating the capabilities of their tools and ensure that they have in place processes and habits to combat any assumptions (individual or institutional) that adding AI to a process will intrinsically bring perfection or solve a problem without potentially creating others. Simply put, no technology is perfect. Therefore, a prospective purchaser/implementer of an AI/ML system should communicate with the vendor about the assumptions/limitations/capabilities of the system to ensure that the system's true characteristics are understood prior to any implementation.

During interviews conducted during the development of this white paper, several participants pointed to technologies that advertised providing AI or ML solutions for the industry even though the tools in fact used other techniques (e.g., simpler statistical techniques). In other cases, active AI/ML systems were not recognized by some participants as AI/ML systems.

Implications to Workforces

Workers across the energy industry, including in the control room, report being overwhelmed and increasingly distracted (e.g., Halverson & Szczepanski, 2019). As operators are being asked to do more work on shift, the industry struggles with an aging workforce, and attempts are made to bring new operators on board faster, the dynamics of the industry continue to change. The growth of AI/ML technologies in real-time operations may help manage many of these concerns, increasing operators' situational awareness and, therefore, reliability. However, if these tools increase operator productivity, resulting management decisions may add additional responsibilities or tasks to the system operator. This would increase operator cognitive load, decreasing the time they have available to review and interact with AI/ML system results and recommendations.

Antifragility

Humans paired with AI/ML systems should focus not only on the current set of circumstances but also on the potential problems that could occur and what to do about them. System reliability and resilience will be bolstered by the human/AI team focusing on potential futures and identifying potential risks and mitigations. AI/ML can present two opportunities here: first, by reducing the amount of cognitive load that an operator faces so that they can spend more mental energy on thinking ahead and, second, by building systems that could be forward-looking.

Implementation Level

Bias in Training Models

Many examples of bias occur due to training data that is incomplete or limited. For example, a generative AI tool that produces text, such as a large language model (LLM), is mostly pre-trained. Techniques with LLMs allow for some updated training to be performed through additional prompts of questions and correct answers, but, generally, the majority of training is done beforehand. Depending on the content that was included in the training/retrieval augmented generation (RAG) phases, the tools may be able to function well in one specific domain/area but not necessarily in others. For example, an LLM trained on one Reliability Coordinator's (RC) operating protocols will likely not function effectively for generating regulatory reports. Users of these tools should be aware of what kinds of training data went into their models' training and what the limits are to their generalizability. Furthermore, when these models are tested and validated, the scope of those tests should be considered as the range of functions in which the tool should be used.

Lack of Generalizability/Agility to New Data or Scenarios

AI solutions often fail as the data evolves over time or when the solution is presented with new situations. These systems should provide users with warnings when input datasets do not match typical patterns on which the systems were trained, informing the users of the possibility of bias. This may be particularly challenging, as many of the confidence intervals provided by these predictive models are still assuming that the data being processed is of the same types and distributions as the data upon which the model was trained.

Recognition of Acceptable Input Ranges

AI/ML systems should be designed to identify appropriate ranges of function (based on inputs and the fit of current data to the training model); when they are outside of that range, those systems should defer to a human if they provide automation, and, for all systems, provide a user with the understanding of the systems' confidence intervals. This is true in aviation, trading automation, and vehicle automation and must be here as well. In order to ensure that a human is ready to supplement or temporarily take over the AI/ML system function, the operator must not fall "out of the loop” and maintain a strong situational awareness. When AI/ML systems, whether purchased from vendors or developed in-house, clearly state these boundaries, it ensures better predictability of system function.

One example of this is the addition of protective measures in the stock market using “circuit breakers” designed to prevent trades from occurring at unreasonable prices, done in response to the Dow Jones Industrial Average dropping nearly 23% in one day in 1987 (Kim & Yang, 2004). While these types of protective measures being activated are certainly not ideal, they prevent further significant degradation of trust in and performance of the system.

Anticipation of System Failures/”Trust but Verify”

Every AI/ML system (like any other technological or human system) should be expected to be imperfect and occasionally fail. These failures can be accidental, due to misconfigurations/errors, or intentionally caused by malicious actors. When humans contemplate technological failures, they can build patterns to recognize risk and understand how to respond better in advance. This reduces the degree of just-in-time (JIT) performance, where users have to think through complex factors in high-stakes, limited-time situations. This forethought is a critical part of antifragility.

When using AI/ML technologies, real-time operators should be able to review, question, and ultimately confirm the accuracy of the system’s results when using the system. This includes the “Explainable AI” problem and systems that can receive feedback from the user (as part of the basic compact). Users should be able to provide feedback often, although they may be less able to do so during busy situations when affected by distraction, fatigue, or other factors, as operators must strive to balance many different concerns in high-stress, high-stakes situations.

Ensuring Staffing Skills for Development, Implementation, and Support

Just as any other technology being integrated into real-time operations, AI/ML tools may require specific skills for support activities. Whether third-party support providers or in-house solutions are used, ensuring that the appropriate expertise is hired and applied is necessary to ensuring that these systems operate correctly, in the same way that other domain expertise and/or certifications are used for other areas (e.g., SCADA, cybersecurity).

Human-System Interaction Level

Cognitive Biases

Because of how our brains work, humans—predictably but not rationally—tend toward certain thought processes in certain situations. Broadly, these kinds of processes, known as cognitive biases, affect our decision making across a variety of situations. While many cognitive biases affect human decision making in real-time operations, a sampling of some that have been observed or are seen as likely to impact successful implementation of AI/ML systems are included here.

Automation Bias

Most AI/ML systems, especially if configured well, are able to function beyond human capabilities in terms of data processing. For example, a human is limited to 7 ± 2 “chunks” of information in working memory, while a computer system can have trillions of bits of data. Humans tend to over-rely on automated systems when there is contradictory information or suggestions that something is wrong (Cummings, 2004). This may be thought of as an assumption by a human that technology has all of the needed variables and proper weightings for all situations, which is rarely the case.

Automaticity

When humans have to focus their attention on something, learn something new, or attempt to contemplate complex problems, our brains use significant amounts of glucose, our primary fuel source (e.g., Legatt, 2015). Because glucose is limited, we have evolved to optimize and conserve resources by identifying and repeating patterns. This allows us to handle a familiar situation without a significant expenditure of resources and in nearly automated modes (e.g., one can safely drive from home to work without remembering the car ride if it was uneventful).

These kinds of behaviors can lead to automatic responses to predictable, consistent situations. For example, if an operator were to receive a recommendation from a tool that they accept (e.g., clicking a “Proceed” button to engage in the recommended action), they are likely to be detail-focused in the first few times using the tool, investigating the recommendation, and ideally reviewing the explanation provided by the tool for why it made the recommendation and what it anticipated would result. Over time, if the tool consistently provides suggestions that are vetted by the human and seen as correct, accurate, and quick, the human could form a habit of “trusting but verifying” less and shifting to “trusting and pressing the button” to quickly pressing the button. This process is likely to happen slowly over time but can lead to a lack of sufficient scrutiny and biases, errors, or incorrect input data in the AI leading to an unintended event.

However, strategies that over-focus on forcing extreme attention (and sometimes, anxiety) onto operators, such as “System I/System II” (Kahneman, 2011), tend to backfire, as they can also bias decision making or overly shift blame onto an individual operator at the point of decision. At the organizational level, this is particularly noteworthy when operators, through the introduction of a new tool that is supposed to improve their efficiency, are expected to “trust and verify” data at higher levels of scrutiny (increasing their cognitive load) while being handed additional operational requirements (decreasing their cognitive resources), potentially preventing them from producing the intended outcomes.

Stovepiping

When receiving data and information, we tend to not only look at the quality of the content but also the source of that data and our levels of trust in it. At times, we tend to “stovepipe,” choosing particular sources of information over all others, especially when contradictory signals are being received. In very much the same way, it is possible for a trusted AI/ML system to provide data to an operator that contradicts assertions from other sources. In those situations, especially with high levels of complexity and/or ambiguity, the operator might focus only on the AI data, ignoring the other signals.

Out of the Loop Syndrome

Understandably, designers of new technologies are focused on technical capabilities, accuracies, and reliabilities. They have typically not yet fully contemplated the relationship between the technology and the humans and the roles that the humans will play, especially when a technological failure or “out of bounds” event requires the humans to take an even more active role. Unfortunately, this also means that new technologies are likely to communicate less with humans and not convey needed data to support their situational awareness. This is often a source of errors in early technology implementations (Endsley & Kiris, 1995).

Within real-time operations, the concept of maintaining reliability through a tool failure (e.g., state estimator) already exists, and this concept should be extended to AI/ML systems, though the question would be slightly different: “If this tool were to fail spontaneously, do I understand the state of the system well enough to solve problems or keep it operating reliably for a period of time?” This situational awareness should not simply involve a reliance on existing tools (e.g., situational awareness tools) but also an understanding of what these tools have been doing/recognizing over time.

Development and Maintenance of Trust

As AI/ML systems can be developed to continuously learn and improve, methods of training humans on their use must be different than more traditional, deterministic systems. For example, an operator must understand that their reliance on a recommendation today may be different (meaning that it should be treated with more or less confidence) than in the past because of other changes on the system. This is more akin to humans working together, as we tend to continually observe others’ behavior to update our confidence intervals in their behavior.

Teaming in a Joint Cognitive System

For humans and technology to work together effectively, both sides must be able to interact, receive and provide feedback, and continue to improve. This concept of building a joint cognitive system, the effective thinking that comes from humans and technology working together, requires several aspects to work effectively (Klein, et al., 2004), as follows:

- Development of "common ground" with common beliefs, understanding, and communication (meaning that the humans and AI/ML will communicate back and forth in ways that both are able to understand)
- Having the ability to predict what the other is likely to do, or why
- Having the ability to give and receive instructions/feedback on how to improve
- Having a shared agreement on how work together will go

Legal Level

Copyright Violations

There are many current examples of companies using vast amounts of public, copyrighted data to train their generative models to produce content, such as for text and image generation. While the courts have not yet definitively ruled on the matter, this is especially important for organizations such as utilities that have significant concerns about risk exposure. A product's use of copyrighted material can expose the user of the product, not just the developer, to significant risks. For example, an AI/ML system could be trained on copyrighted material then used by an operator to write a DOE-417 report or generate slides that get used by an employee for a public presentation. In such a case, this can unintentionally expose the utility to liability.

For a system operator using a generative tool to generate reports, concerns about these kinds of risks may lead the operator to focus on certain concerns (e.g., rephrasing certain text) and shift their focus away from other concerns (e.g., ensuring that a conclusion is valid).

Informational Leakage

During informal interviews and in survey results, significant concerns were raised around ensuring the security of private or confidential information, especially around Critical Energy Infrastructure Information (CEII), whether being sent to or returned by generative AI systems. Already, operators who write reports have to consider what information to include or exclude based on reporting and confidentiality requirements. Leveraging generative AI to write these kinds of reports may lead to occasional inclusion of information that an operator would not have wanted in the document. Depending on how the generative AI was trained, for example, it may be unable to distinguish between two different report types, one that must include certain CEII and another that must not.

Allocation of Risk

Across many domains where AI/ML is being used, there are significant legal concerns around the allocation of risk. Autonomous vehicle systems are a strong example of this, as the courts have not yet determined what happens when an autonomous vehicle makes a decision that leads to an accident or a violation of the law. Similarly, as real-time operations look at a wide range of use cases for AI/ML technologies to replace and/or supplement human decision making, there may be questions about what happens when the technology provides an incorrect answer.

For example, when a technology that uses AI/ML to perform load forecasting (e.g., "similar day" analysis) picks incorrectly and leads to significant under-commitment of generation, what occurs next? If the operator was unable to understand why that decision was made (due to poor explainability), does that factor in? This paper does not focus on suggesting answers to these kinds of legal questions but rather seeks to indicate that ambiguities that currently exist can create fear, uncertainty, and doubt in the minds of people in real-time system operations, potentially drawing attention and vigilance away from critical work or increasing the likelihood of undesirable outcomes.

Production Strategy and Version Controls

As AI/ML systems receive data and make predictions or recommendations that affect the reliability of the BPS, there are several additional important considerations, both in terms of operators' capacity to trust these systems and in legal terms. These include the following:

- Version control: Ensuring that changes to the underlying programming of AI/ML systems, or updates to their training models, are properly tracked over time
- Change management: Ensuring that installations and modifications to the computer systems running AI/ML systems are systematically tracked and documented
- Data provenance: Ensuring that the path of the data is documented through its sourcing, collection, and transformation as it moved into the AI/ML system
- Data quality: Ensuring that the data's accuracy, completeness, and consistency is tracked, including data in both training and live usages
- Data lineage: Ensuring that the flow of data through the AI/ML system is well documented and understood

Recommendations

As entities across the industry identify increasing complexity and complicatedness in their real-time operations, AI/ML technologies are posed to significantly help operators maintain the reliability of the BPS if implemented well. As such, several recommendations are made to ensure that these technologies and the humans who work with them are able to more effectively work together while lowering the risk of error.

In many ways, AI/ML systems are different from other traditional computer systems in how they arrive at answers, interface with users, and—in some cases—self-update over time. Furthermore, because of their potential roles in helping operators respond quickly and effectively to increased system complexity, they are likely to shift how an operator would do their work on shift and what is needed of them to maintain system reliability.

Human/Organizational Performance

A significant amount of research, science, and training is available to support organizations in human/organizational performance training. In 2025, the Department of Energy will release its updated Human and Organizational Performance Improvement Handbook, the successor to the Human Performance Improvement Handbook of 2009. This and many other energy-specific resources are available to address many of the human factors above and provide tools and techniques to reduce the risks of human errors, both in real-time operations and management. Many organizations have robust groups to address human and organizational performance, including training, safety, and dedicated HOP groups. Many use tools like root-cause analysis, learning teams, and near-miss reporting to learn and improve performance following an event and to reduce the likelihood of future events.

Training for AI/ML System Use

More than ever, training provides a critical precursor to real-time operations, ensuring that operators have the needed mental models, habits, and skills before entering the real-time control arena with AI teammates.

Some examples of recommended skills to be covered in training include the following:

- For systems that are self-updating/training over time, an understanding that past reliability may be different than current reliability
- For systems that are dealing with increasing amounts of data or abstraction, the understanding that the way that an answer is arrived at may be different than expected (explainable AI problem)
- An understanding of the appropriate levels of "trust but verify" in the tool's operation

- Ensuring the monitoring and review of the system's predictions/conclusions over time (an increasingly important part of situational awareness)
- When using generative technologies:
    - Ensuring critical review of content before submission, such as identifying and addressing ambiguities, and ensuring no confidential information is included
    - Understanding downstream uses of submissions and how they will be processed (e.g., a GPT is used by one entity to generate a report, and another entity uses a different GPT to summarize it)
    - Validating the content and ensuring that no inappropriate assumptions/conclusions/hallucinations are present in the content (e.g., Schiller, 2024).

Training for Wide-Area Knowledge and Expertise

Over the past several years, other changes to the system (e.g., growth of renewables and a "silver tsunami" of retiring workers) have spurred a need to onboard new operators quickly and effectively. As in many other industries, historically longer training times with many different on-site focuses have shifted into more rapid and easily assessable functional testing. Today, a new system operator may have sat through many engineering classes and obtained system operator certifications and not ever have been on site at a generator or substation.

However, some of that on-site knowledge may become increasingly important again, as operators will be increasingly responsible for understanding, validating, and approving conclusions reached by AI/ML systems. Therefore, ensuring that operators build and continue to maintain wider-area knowledge and expertise about the state and complexities of the system will better prepare them to understand the broad implications of the recommendations from AI/ML systems.

Training with High-Fidelity Simulations

The use of high-fidelity simulations is among the key ways that enhanced training may better prepare operators for human-AI teamwork. Although traditional aspects of training (e.g., questions on procedures or requirements or tabletop exercises and "what if" scenario discussions) certainly still hold value, the most effective way to ensure the necessary skills is through the use of these simulators since operators must be able to better interact with, and predict behaviors of, their core systems. These kinds of advanced simulations build core skills, mental models, and habits, which can then flow into real-time operations.

This approach also has the added benefit of improved evaluation capabilities and better knowledge/habit gap analyses. High-fidelity simulators are a much lower-risk approach to learning how to quickly understand and interface with these systems. However, for these simulations to be effective, they must mimic as closely as possible the real-world systems and scenarios that the operators will be facing. For example, a simulator that provides simulated relay trips for transmission line overcurrent but ignores under/over-frequency relay action for units could potentially introduce gaps in system operator training, which could ultimately lead to an undesirable event in a high-stress, high-stakes situation.

Chapter 4: Industry Survey Results

During the generation of this white paper, several industry subject matter experts, including members of several NERC working groups and committees, responded to a survey on their understanding, trust, and use of AI/ML technologies. During this period, (n=47) responses were received.

Demographics

Figure 4.1 shows survey participants crossed a wide range of the ERO Enterprise, from the federal, NERC, Balancing Authority (BA), RC, and Regional Entity levels to transmission, distribution, and generation providers. Those marked as "other" included national labs and trainers. Approximately 40% of the participants held manager/director positions, 11% were executives, and 30% worked in real-time Operational Technology (OT) operations and/or Security Operations Center (SOC)/Network Operations Center (NOC) cyber operations.

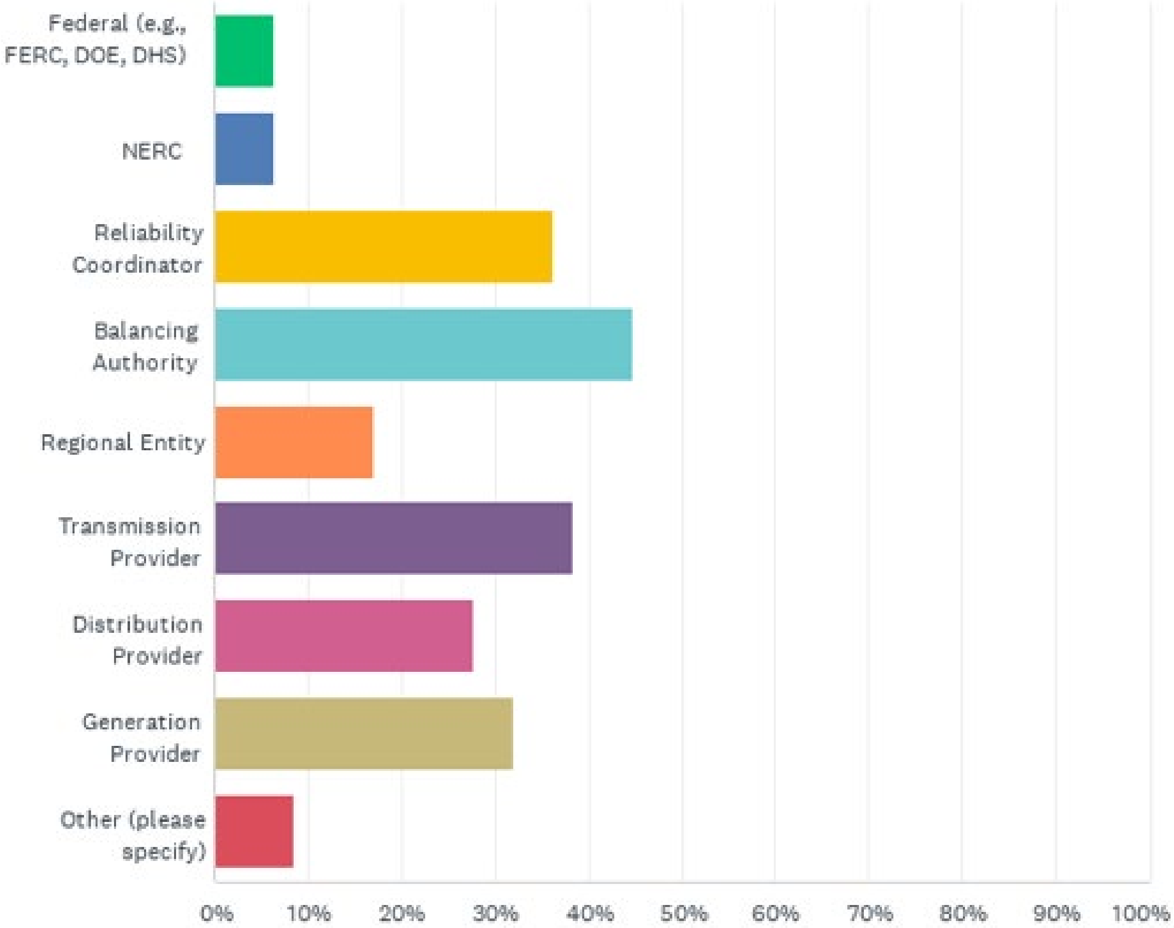


Figure 4.1: Q1 What Kind of Organization Do Your Work For?

Current Levels of Knowledge/Expertise

Figure 4.2 shows that participants were requested to self-categorize their level of knowledge and involvement with AI/ML technologies. Most (47%) of the participants considered themselves to be learning about them but not directly having implemented or built them. Some (9%) had built/configured systems, including generative models, while others had implemented existing AI/ML tools (17%) and others reported using existing tools (15%). Another 15% reported knowing little about AI/ML technologies and not being aware of using them.

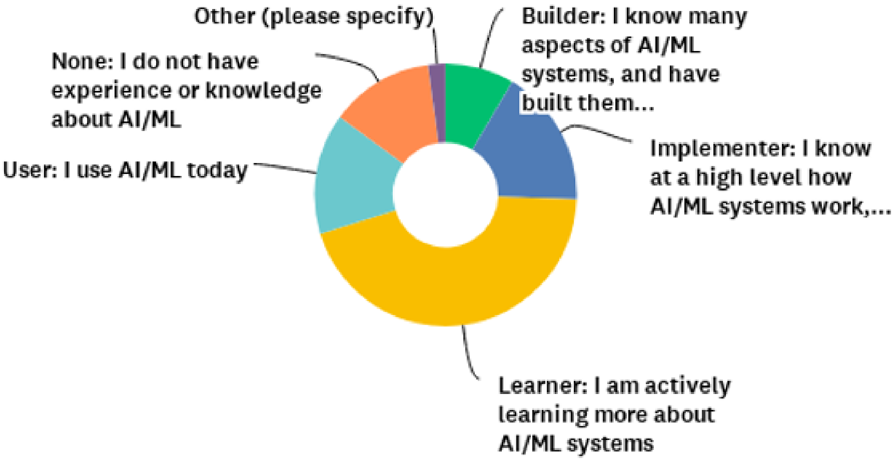


Figure 4.2: How Would You Describe Your Knowledge of AI/ML?

Organizational Approaches to AI/ML

Participants were asked to identify which approaches their organizations were taking to AI/ML technologies. As with the informal interviews, it was noted that these organizations had vastly different approaches on how to view/implement/ban the technologies. During informal interviews, some participants identified products that used some form of AI/ML technologies (e.g., a wind forecasting tool) but were not aware it was AI/ML based, and, in other cases, had thought a tool was using AI/ML when it was not. Some examples were provided of marketing materials for those tools that also overstated their use of AI/ML technologies (e.g., performed simpler statistical analyses).

Table 4.1: Identify Approaches Organizations Currently Taking to AI/ML Technologies

Approach	Percent
We have guidelines/policies in place around the use of AI/ML, from cybersecurity and information security perspectives	53%
We use AI/ML tools for the prediction of future events (e.g., weather, prices)	38%
We have active bans on using public AI tools (e.g., ChatGPT, Microsoft Co-Pilot)	30%
We have internal AI/ML tools we can use in-house or on a private cloud (e.g., our own large language model systems)	26%
We use AI/ML tools to help with cybersecurity	19%
We use AI/ML tools for predicting equipment failures	11%
I’m not aware of any AI/ML use within the company	8%
We are starting to define our approach to AI/ML (e.g., generating use cases)	6%
We use AI/ML tools to help with customer service	4%
We use AI/ML tools to help with compliance, regulatory, or other reporting activities	2%
We use AI/ML tools to help with physical security	2%

Perspectives on AI/ML

Participants were generally positive about the value and potential of AI/ML technologies. Some concerns were expressed about the risks of using AI/ML technologies and growing ethical, legal, and operational concerns in the future.

Table 4.2: Value and Potential of AI/ML Technologies

Statement	Strongly Disagree	Disagree	Agree	Strongly Agree
They are the future		6%	38%	56%
They can do work better than humans can in some ways		6%	66%	28%
They are helpful when configured/used properly			47%	53%
They're not perfect and need humans to keep them working properly		3%	31%	66%
They are very interesting			50%	50%
They are a trending thing now	3%	6%	38%	53%
Over time, they won't be as major a factor in our lives	41%	53%	6%	
They are impressive today	9%	16%	59%	16%
They continue to grow over time			50%	50%
There are lots of legal and ethical questions we need to answer as we rely on these systems more and more		6%	22%	72%
I find them scary	28%	38%	28%	6%
I find them hard to understand	19%	41%	38%	3%
I'm afraid it will put me out of a job	44%	44%	9%	3%
I don't know what I'll need to do if the system doesn't work properly	16%	56%	22%	6%
I will know when it's not working properly or giving me a bad answer	3%	44%	47%	6%
They can help me do my job better in high-stakes situations	3%	13%	65%	19%
They can help me do my day-to-day job better	3%	6%	65%	26%
They are incredibly risky	13%	47%	38%	3%
They should be avoided at all costs	59%	38%		3%

AI/ML Use Cases

Across several use cases, participants were asked to rate their organization's current stance on AI/ML tools across the following criteria. For display purposes, these were coded from 0 (Will Not Use) to 5 (Currently in Production):

- 0 - Will Not Use
- 1 - Haven't Considered
- 2 - Currently Considering
- 3 - Currently Testing
- 4 - Currently Implemented, Validating
- 5 - Currently in Production

AI/ML Application by Use Case

Table 4.3 shows that, when analyzed by use case, AI/ML use varies across the industry, with particular focus on prediction of system load/demand and renewable generation levels. These use cases are shown, sorted by averages for those activities:

Table 4.3: AI/ML Use-Case Analysis

Activity	Count	Mean	Std Dev	Min	25%	50%	75%	Max
Predicting system load/demand	41	2.83	1.63	0	2	2	5	5
Predicting renewable generation	41	2.20	1.36	0	1	2	3	5
Cybersecurity-related	42	2.12	1.21	1	1	2	3	5
Predicting temperature, humidity, etc.	40	2.00	1.43	0	1	2	3	5
Predicting equipment failures	41	1.76	1.16	0	1	2	2	5
PMU/Synchrophasor Monitoring	40	1.60	1.10	0	1	1	2	5
Predicting unit availability	38	1.55	0.98	0	1	2	2	5
Improving customer service	41	1.51	0.81	0	1	1	2	4
Predicting transient systems stability issues	40	1.50	0.75	0	1	1.5	2	3
Predicting prices	39	1.49	0.85	0	1	2	2	4
Improving state estimation	40	1.48	0.88	0	1	1	2	4
Writing compliance reports	41	1.41	0.81	0	1	1	2	4
Writing log entries	40	1.35	0.77	0	1	1	2	4
Predicting wildfires, storms, etc.	41	1.29	0.84	0	1	1	2	4
Predicting fire risks	40	1.18	0.81	0	1	1	1.25	4
Generating switching orders	40	1.08	0.66	0	1	1	1	3

AI/ML Application by Entity Type

When analyzed by self-identified group membership, the average responses for each of the questions and criteria are shown below. There was significant variance in the degree of testing and implementation of these technologies, as different organizations within the industry focused on different aspects of AI/ML technologies. Several organizations have already implemented AI/ML tools in real-time operations.

Table 4.4: AI/ML Use Case Analysis

Role	Count	Mean	Std Dev	Min	25%	50%	75%	Max
Reliability Coordinator (RC)	259	1.69	1.17	0	1	1	2	5
Balancing Authority (BA)	323	1.74	1.10	0	1	2	2	5
Regional Entity (RE)	79	1.49	1.07	0	1	1	2	4
Transmission Provider	286	1.74	0.98	0	1	2	2	5
Distribution Provider	206	1.70	1.01	0	1	1	2	5
Generation Provider	238	1.74	0.96	0	1	2	2	5

Free-Form Perspectives

In free-form responses, participants varied in their perceptions about AI and their concerns. Overall, the following sentiments were noted repeatedly:

- Concerns about cybersecurity
- Concerns about biases in model training
- Concerns about release of private/confidential/CEII information into the public
- Recognition of the need for forecasting in generation, especially about intermittent renewables, and fuel supply
- Recognition that malicious actors can use the same tools used by system operators to harm the system
- Recognition that the way we process and use information in real-time system operations is changing (as one participant put it, “We are moving from the information age to the intelligence age”).
- Recognition that AI/ML tools can significantly improve productivity, efficiency, and effectiveness
- Recognition that society is currently heavily focused on AI/ML technologies but with limited experiences using the technologies over time
  - One participant wrote, “It would be good to submit this survey one year from now after the hype cycle turns, to see what’s changed.”
  - Another participant wrote, “There are a lot of unknowns and assumptions about AI/ML to have clear answers now.”
  - Another participant wrote, “I would suggest that some of what is perceived as advanced is really just inscrutable and not independently verifiable…and that is part of the problem.”

Chapter 5: AI/ML in System Operations

AI/ML technologies are being investigated or are in use for a variety of use cases in real-time system operations. Many organizations are looking at using AI/ML to perform existing functions and meet requirements. This section chronicles some examples of active development or use of AI/ML across the industry. It should not be considered an exhaustive list of possibilities for future use.

Load Forecasting

MISO Load Forecast Modeling

MISO's load forecasting process (**Figure 5.1**) consists of short- and long-term models. Both models share similar inputs and begin with examining individual local BA (LBA) profiles, including historical load and weather data and heating and cooling degree days. Each LBA's load forecasting model creates and uses a base model with regional awareness. The individual models are personalized by incorporating real-time and historical data and finetuning them accordingly.

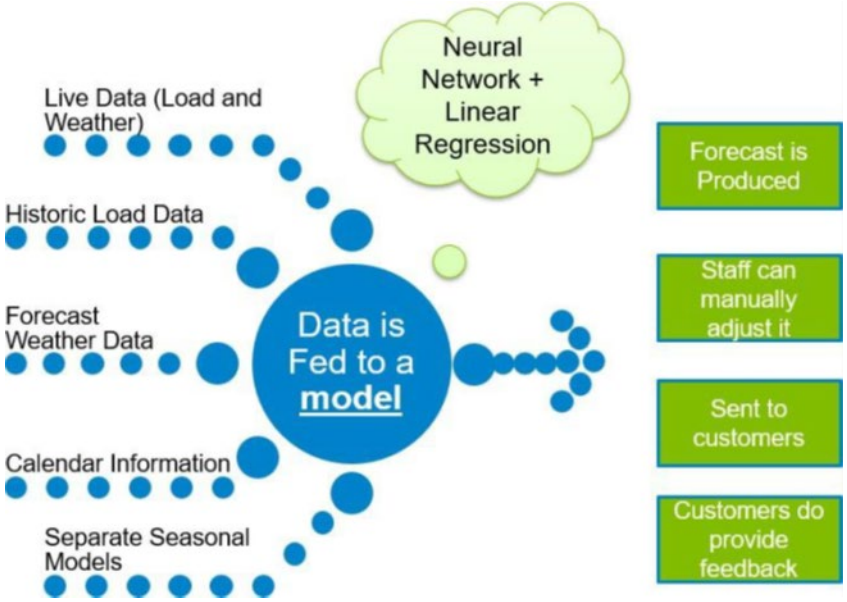


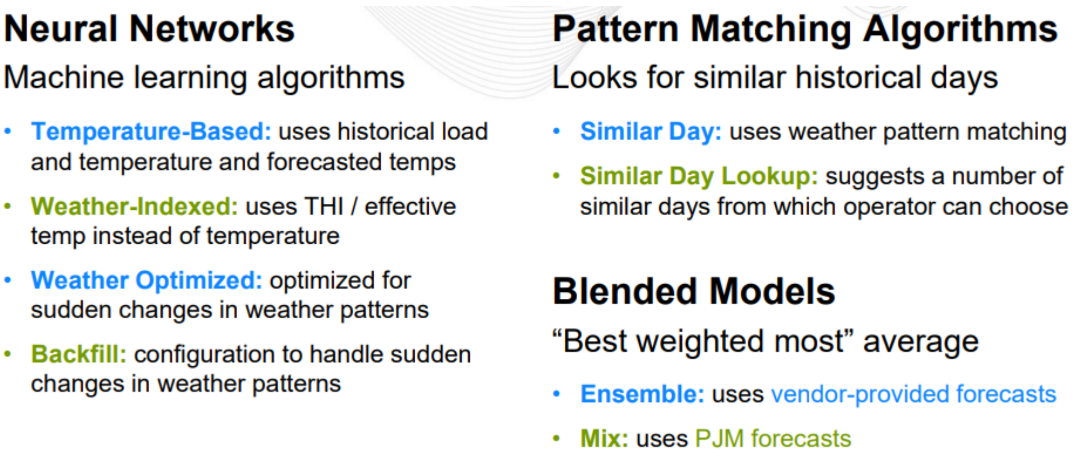
Figure 5.1: MISO Load Forecasting Model

Neural network and linear regression software allows for continuous learning and adjustment of the models with updated information. The short-term model relies on recent forecasts and actual load values. In contrast, the long-term model uses load and weather data for accurate updates. If an LBA has zero load for a period, no load forecast model is developed, and the load is reported as zero in the aggregate MISO market load (MISO, 2023).

MISO has adopted the use of linear regression for short-term load forecast (STLF) and forecasts the load over a five-minute interval. Mid-term load forecast (MTLF) utilizes neural networks to create a seven-day forecast and is typically used in real-time and forward operations. Both forecasts' predictions have been active since 2013. The MISO operations forecast team does monitor the forecast for anomalies and will start an ad-hoc analysis if a weather risk is identified outside the inputs to the process.

PJM Load Forecasting

Similarly, PJM uses both ML and pattern matching algorithms in its load forecasting practices (**Figure 5.2**). System operators are presented with a blend of these models based upon predetermined weighting developed by engineers and forecasting experts, along with the ability to use their own experience to further refine the forecasts.



Neural Networks

Machine learning algorithms

- **Temperature-Based:** uses historical load and temperature and forecasted temps
- **Weather-Indexed:** uses THI / effective temp instead of temperature
- **Weather Optimized:** optimized for sudden changes in weather patterns
- **Backfill:** configuration to handle sudden changes in weather patterns

Pattern Matching Algorithms

Looks for similar historical days

- **Similar Day:** uses weather pattern matching
- **Similar Day Lookup:** suggests a number of similar days from which operator can choose

Blended Models

"Best weighted most" average

- **Ensemble:** uses vendor-provided forecasts
- **Mix:** uses PJM forecasts

Figure 5.2: PJM Load Forecasting Practices

Many refer to current load forecasting practices as both "an art and a science." The "science" refers to the algorithms, which may include AI/ML. The "art" comes from the need to understand and predict how load will change based on circumstances, such as holidays, major sporting events (e.g., Super Bowl, Olympics), and extreme weather. Traditional ML models have struggled to provide reliable results in these circumstances. In today's world, humans fill this "art" role, but it nonetheless remains a challenge.

As we move into the next phase of the computing revolution and the expansion of AI, this "art" is an area where we may be able to further expand the science and improve the forecasting on a daily basis, but the improved forecasting for extreme days is of significant value. Doing a better job of predicting as opposed to trending will improve both the reliability and efficiency of grid operations. Forecast subject matter experts should continue to evaluate the state of their current technology and look for opportunities to work across the industry with peers to collaborate in implementing AI and developing the next generation of forecasting models.

Solar and Wind Forecasting

As the industry continues the energy transition from traditional fossil fuels to renewable sources of energy, control room operators and operations engineers are now tasked with a job that is exponentially more complex than it has been for the past 100 years when it comes to planning and operating the grid. The task was a challenge when forecasting the electric demand was the biggest uncertainty, but there are now increased uncertainties that arise when forecasts for solar and wind generation must also be considered when it comes to ensuring that the accurate balance of generation to load will be maintained. Further complicating this are the behind-the-meter installations of solar and other distributed resources whose output is affected by both weather conditions and other human factors.

The current state of the art is for grid operators to use existing ML and weather models to develop forecasts for load, as noted above, and layer on top the forecasts for wind and solar to develop a "net load" for which traditional generation will need to be dispatched to follow. In addition to the challenge of forecasting each of these three major components of uncertainty accurately, there are challenges in predicting the range of errors inherent to each forecast to then informing the operating plan for how many MWs to prepare to dispatch and how many MWs may need to be held in reserve. When there are larger uncertainties, larger MWs may need to be held in reserve to then be able to respond if those uncertainties materialize. This is a prudent and reliable way to operate but leaves room for improvement and efficiencies to be gained.

This is where AI may be a useful application. Similar to applications to improve load forecasting, AI models may be able to provide significant improvements to both the accuracy of the models as well as rapidly identifying times of greater risks when the solar and/or wind forecasts may be more volatile. Examples include identifying periods of potential high wind speed cut off, providing improved models for icing conditions, and predicting when “snow slip” (when snow that has accumulated on solar panels slips off due to warming conditions that then results in a sudden jump in solar energy production) may occur. Every incremental improvement to increasing the accuracy of these forecasts translates to an incremental improvement to both reliability and the efficient operation of the grid, both of which are ultimately to the benefit of the end-use customers. Exploring these opportunities using AI is the next step on which industry should be collaborating with the National Labs and the computing industry.

Contingency/Stability Analysis

During times of unplanned and unexpected system events that may result in the loss of multiple transmission elements and generation resources (e.g., tornado, derecho, attack), the system operators must react as quickly as possible to assess and secure the systems. Operator training, robust operating practices, support staff, emergency planning, and advanced control room tools all contribute to ensure this happens today.

AI may have a role among the control room tools that help set the system operators up for success in maintaining reliability. As previously noted, ML and AI have two major applications: analyzing large quantities of data to draw conclusions and processing high volumes of written text and responding to queries with summaries and/or citations of said text. During system events where alarms are flooding the operator with information that they need to quickly process and respond to, there could be enhancement by specific implementations of new AI-based applications. For example, if the AI engine was able to help process the alarms, summarize key events, and recommend procedures to apply, this would serve as a welcome aid to operators to help expedite the triage and response process.

Outage Management

In recent years, ML has been used to help predict when and where equipment failures may occur. Such practices are then used to perform preventive maintenance to avoid unplanned failures and are also used to streamline and enhance the maintenance programs that generation owners and operators employ. Advances in AI may be able to further improve and expand upon the use of such practices.

Report Generation and Procedure Drafting

In today’s paradigm, the status quo for event analysis consists of engineers spending days to weeks gathering post-event data, evaluating metrics, and performing analysis to start putting together the picture of what occurred. The next step involves taking this information and distilling it down into a clear narrative with a consistent voice. This is where additional people are needed with a background in report writing, proofing, and grammar checking. Depending upon the event being reported on, this process can take months.

The introduction of AI and LLMs, such as OpenAI ChatGPT, Google Gemini, and Meta Llama, has ushered in a new era in possibilities for the application of this technology. Report writing may be one of them. It is now possible to input thousands of pages of written information into these models and within seconds have a clear summary of the information provided to the user. By extension, the information gathered by the engineers, or even just the raw data itself, could be fed into an LLM to then draft the report, or at least provide a robust and well-developed framework for a report, as opposed to starting from scratch. Theoretically, this could speed up the current process considerably to just a few hours or days to review, edit, and publish the report. More importantly, this would allow engineers and system operators to better spend their time on reviewing the results, developing action items/recommendations, and implementing the changes.

AI may also be able to assist with producing clear, concise, and grammatically consistent procedure writing. As with report writing, it is not often that the experts with the system knowledge needed to write procedures also have a background in technical writing. This is an area where the LLMs may be able to take the technical details and quickly draft clear and accurate procedures. This would have the benefits of reducing subject matter expert time spent drafting the procedures and providing clear, concise, and efficient procedures for the operators to use.

EMS and Planning Model Validations

Accurate electric system models are critical to the reliable planning and operation of the grid. Engineers and model experts spend countless hours validating these models, updating them for future configuration (new equipment outages, etc.), and coordinating and aligning them between companies. A host of tools have been developed to assist in these validations, but these models do contain a vast amount of information, and the validations become exponentially more complex when real-time telemetry and historical system/event response data is introduced. This is an area where utilization of AI to analyze these vast datasets and provide new insights into model anomalies may be highly beneficial. Furthermore, AI may be able to go beyond identification of anomalies and actually recommend or correct the models to address those anomalies.

System Operator Training

The power industry, like other critical industries that serve the public needs, must have the best-trained operators to prepare for and respond to any system event. A key element to being successful in this endeavor is training. At the current state, most training is performed using off-line simulations combined with self-study and real-time shadowing, whereby the trainee is paired with a qualified system operator/engineer. This model works very well and is evidenced by the overall reliable operation of the grid to date.

However, a limitation of the current training model is the ability to expose the trainee to a wide variety of unexpected situations. Typically, the training only covers unusual events if they have already occurred. In a world where extreme weather, for example, is happening more often, AI may offer opportunities to train the operators on situations that have not yet occurred so they can be better prepared for how to respond. Additional aspects of training are covered in **Chapter 3, Recommendations**.

Anomaly Detection

AI/ML systems that can make predictions/perform pattern recognition can also be used to identify violations of those expected patterns. For example, a system that is trained to identify normal ranges/patterns of telemetry, or typical behaviors on equipment or people, could also be used to identify out-of-range behaviors. Assuming these systems are designed to present information to operators clearly, they can help identify equipment failures, malicious actors, and other undesirable situations, which in turn could alter the operator’s behavior.

Chapter 6: Cybersecurity Issues Related to AI/ML Technologies

AI and ML technologies are a boon to the development of applications used for critical infrastructure. However, they can also be used to highlight deficiencies in operating environments. In the realm of cybersecurity, some of these deficiencies can be exploited by threat actors to gain access to, and even take control of, environments. Just as with defensive measures, adversarial activities can be enhanced and—in some cases—enabled by AI/ML technologies and tools.

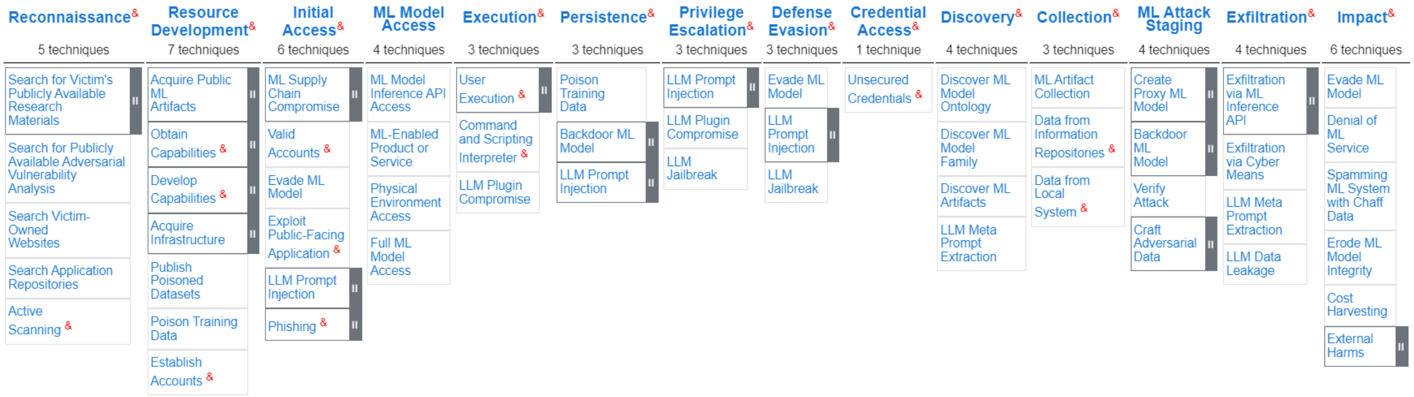
Risk Management

Risk is typically viewed in relation to threats, assets, and vulnerabilities. As pointed out by Adam Shostack in his classic book Threat Modeling: Designing for Security (Shostack, 2014), there exists an interplay between requirements, threats, and mitigations in a system. Shostack's STRIDE Model involves: Spoofing identity, Tampering with data, Repudiation threats, Information disclosure, Denial of service, and Elevation of privileges. Viewed as an application, an AI system must also be hardened to meet the consequences of the varying aspects of risk management. The Artificial Intelligence Risk Management Framework (NIST, 2023) developed by the AI Working Group and the National Institute of Standards and Technology (NIST) can be used to identify unique risks posed by AI while also proposing actions for mitigating those risks. Management of the risks and therefore the trustworthiness of a system is a process of mitigating the risks created by vulnerabilities inherent in the software that comprises the operating system, firmware, application environments, network protocols, storage technologies, and related utilities as well as the potential for systemic access issues through erroneous configuration settings. Likewise, researchers at Georgetown University's Center for Security and Emerging Technology (CSET) indicate that mapping and navigation of "the terrain of AI risk and harm" is essential.

Artificial Intelligence Threat Modeling: Threats, Assets, and Vulnerabilities

Threat matrices have been designed for various types of technology frameworks. In the IT and OT spaces, MITRE's ATT&CK Framework (2024) is well known. An ancillary framework called ATLAS (2024) has been developed for analysis of threats to AI infrastructure. The parallels to the ATT&CK Framework are readily apparent, and the forward-thinking nature of the techniques found in the impacts listed in the rightmost column highlight the potential for adversarial outcomes to the health of the AI itself as well as the organizations that depend on it.

The MITRE ATT&CK Framework is typically laid out in a series of matrices, with each focusing on a specific area from enterprise architecture to industrial controls, cloud, and other types of computing environments. Each matrix is designed with columns representing tactics and rows representing corresponding techniques as shown in Figure 6.1.



Reconnaissance&	Resource Development&	Initial Access&	ML Model Access	Execution&	Persistence&	Privilege Escalation&	Defense Evasion&	Credential Access&	Discovery&	Collection&	ML Attack Staging	Exfiltration&	Impact&
5 techniques	7 techniques	6 techniques	4 techniques	3 techniques	3 techniques	3 techniques	3 techniques	1 technique	4 techniques	3 techniques	4 techniques	4 techniques	6 techniques
Search for Victim's Publicly Available Research Materials	Acquire Public ML Artifacts	ML Supply Chain Compromise	ML Model Inference API Access	User Execution &	Poison Training Data	LLM Prompt Injection	Evade ML Model	Unsecured Credentials &	Discover ML Model Ontology	ML Artifact Collection	Create Proxy ML Model	Exfiltration via ML Inference API	Evade ML Model
Search for Publicly Available Adversarial Vulnerability Analysis	Obtain Capabilities &	Valid Accounts &	ML-Enabled Product or Service	Command and Scripting Interpreter &	Backdoor ML Model	LLM Plugin Compromise	LLM Prompt Injection		Discover ML Model Family	Data from Information Repositories &	Backdoor ML Model	Exfiltration via Cyber Means	Denial of ML Service
Search Victim-Owned Websites	Develop Capabilities &	Evade ML Model	Physical Environment Access	LLM Plugin Compromise	LLM Prompt Injection	LLM Jailbreak	LLM Jailbreak		Discover ML Artifacts	Data from Local System &	Verify Attack	LLM Meta Prompt Extraction	Spamming ML System with Chaff Data
Search Application Repositories	Acquire Infrastructure	Exploit Public-Facing Application &	Full ML Model Access						LLM Meta Prompt Extraction		Craft Adversarial Data	LLM Data Leakage	Erode ML Model Integrity
Active Scanning &	Publish Poisoned Datasets	LLM Prompt Injection											Cost Harvesting
	Poison Training Data	Phishing &											External Harms
	Establish Accounts &												

Figure 6.1: MITRE ATLAS AI Matrix (MITRE Corporation)

The tactics currently laid out in MITRE's AI matrix as shown in **Figure 6.1**, referred to as ATLAS, consist of the following:

- **Reconnaissance**
    - Techniques focusing on searching for available information and actively scanning the victim environment
- **Resource Development**
    - Techniques focusing on developing capabilities and acquiring infrastructure
- **Initial Access**
    - Techniques focusing on gaining access to the victim environment through the use of accounts, applications, and various compromises
- **ML Model Access**
    - Techniques focusing on access to the model and/or the victim's physical environment
- **Execution**
    - Techniques focusing on user execution or the use of compromised software or access
- **Persistence**
    - Techniques focusing on establishing a foothold via a backdoor or by using poisoned data or malicious prompts
- **Privilege Escalation**
    - Techniques focusing on compromise through processes such as malicious prompting, LLM plugins, or "jailbreaking" the model's environment itself
- **Defense Evasion**
    - Techniques focusing on evasion of security measures through malicious prompting or by "jailbreaking" the model's environment
- **Credential Access**
    - Technique(s) focused on accessing unsecured credentials
- **Discovery**
    - Techniques focused on understanding the underlying model through analysis of artifacts, information about the ML model itself, or through extracted prompt metadata
- **Collection**
    - Techniques focusing on collecting ML artifacts by accessing data from the underlying local system or from information repositories
- **ML Attack Staging**
    - Techniques focusing on staging a proxy model or by exploiting a backdoor in the ML model (e.g., verifying an attack method through the abuse of an API, or using crafted adversarial data)
- **Exfiltration**
    - Techniques focusing on unauthorized extraction of data via an API, via meta prompts, or through LLM data leakage

- **Impact**
  - Techniques focusing on affecting the performance of the model (e.g., evading protections, creating a denial-of-service condition, eroding the integrity of the model, wasting compute resources, or causing financial and/or societal harms)

Additionally, the "AI as a Service" model, as outsourced AI engines hosted in cloud environments are being used on an increasing basis, is beginning to highlight additional security exposures. A second model is the MITRE ATT&CK Cloud Matrix (MITRE Corporation, 2024), which highlights the underlying cloud technology used by many AI models.

Various scenarios providing an outline of the threats through which the use (or abuse) of AI-related technologies could create potential impacts to operations have been realized and highlighted by researchers at OpenAI (Goldstein et al, 2023) and could include the injection of tainted data into the model for adversarial purposes, as follows:

- **Poisoning**
  - Injecting tainted data into the training dataset of a model for adversarial purposes
- **Evasion**
  - Altering input to change how the system responds to inputs after the model is deployed
- **Inference**
  - An ML security threat involving using the output of a model to infer its parameters or architecture
- **Prompt Injection**
  - A manipulation of an AI system by feeding malicious inputs disguised as legitimate user prompts
- **Other Established Cyber Attack Methods**
  - Attack methods typically found in software that take advantage of vulnerabilities in underlying software, such as software supply chain attacks, scripting attacks, or secrets such as stored keys or embedded static passwords

Assets

An asset from the perspective of an AI system does not necessarily equate to a physical asset as we have historically considered in IT resourcing. The quantity and quality of data in AI are more ethereal constructs as the data is both used for training the models and for providing content as output. Realistically, there exist physical assets, but they tend to be abstracted by the virtualized environments that are essential to processing the large scales of data required to train meaningful models. Assets consist of intellectual property such as data, models, software, and code.

Data

Access to data residing in internal sources has been used by external entities to train their AI models and has been the subject of prominent lawsuits. Intellectual property rights remain a concern in terms of organizational confidentiality, as externally available corporate and employee information could be used for intelligence gathering and corporate espionage. Training data poisoning could prove insidious as the validity of the output of the system to users cannot be verified. Data integrity and confidentiality are required to ensure that the provenance of the corpus, not only of production data but also of training data, is maintained and protected.

One serious challenge to the output of current AI systems is the inadvertent production of situations where false inferences are created and returned as what are termed hallucinations, which have the potential to create false associations and narratives, resulting in invalid information returned by AI models (typically, but not limited to, LLMs) related to organizations, employees, business partners, or even customers.

There is a parallel to the cloud model concept in that an organization merely leases the computing resources of the cloud infrastructure to train an AI model. The organization may or may not own the AI model, but the responsibility for the security and integrity of the data belongs solely to the custodial organization.

Models

A model is a dataset used to train an AI system on a set of data, or training data, to recognize defined patterns or make defined decisions known as inferences. A model can and will become outdated relative to shifts in the deployed context of the training data due to changes in the relevance of the data over time. Models are available as open-source tooling or can be developed by an entity with data science expertise. Sites such as HuggingFace and GitHub contain repositories related to various open-source models.

Software

Software supply chain vulnerabilities are a critical issue facing cybersecurity professionals and managers. These vulnerabilities, through the development process, have been carried forth into the software supply chain used by AI and ML technologies.

> DevSecOps [is used] to produce code faster and at lower cost, but the reality is that much of the code is actually coming from the software supply chain through code libraries, open source, and third-party components where reuse is rampant.
>
> - Carol Woody, a principal researcher in the Software Engineering Institute's (SEI) CERT Division (Woody, 2022)

There is a relative scarcity of expertise in AI engineering and development as highlighted in the August 2024 *US State of Generative AI in the Enterprise Quarter three report* by Deloitte. The report indicates that 20% of respondents "think they are…highly prepared for GenAI" in terms of available "talent," and therefore the majority of resources used for development and hosting of these environments are acquired through the software supply chain (Deloitte, 2024). Software such as Ray AI, repositories such as the aforementioned HuggingFace and GitHub, and languages and support utilities such as Python, Jupyter Notebooks, PyTorch, Keras, and LangChain include a multitude of utilities, support libraries, and packages hosted in various (sometimes additional) repositories. Unfortunately, as noted by researchers at BishopFox (Garcia, 2023) and amplified by Forbes (Brewster, 2024), examples of vulnerabilities in software and errors in system configurations can allow for data breaches affecting developers of the software as well as their customers.

Collaboration between government and industry stakeholders through threat intelligence monitoring and information sharing can alleviate potential hazards presented by default environment configurations by establishing developer and customer community best practices.

Code

Beyond the software used to design the underlying model and the related environment that will support it exists the code that must be developed by the team(s) responsible for integration and deployment of the AI. As important as the security of the underlying architecture may be, the code developed in the high-level programming or scripting language(s) implemented to bind and support the various components of the system, often referred to as "glue" by developers, must also follow strict development principles and guidelines with the goal of preventing the introduction of vulnerabilities into the system.

Vulnerabilities

A vulnerability in software is typically introduced by a development error, or a flaw in underlying or inherited code in an executable, a library, or a function. There are various mechanisms through which a vulnerability can be introduced into an AI software environment. This list is not exhaustive but highlights issues under review by security researchers.

Hallucinations

A hallucination is the concept of the return by an AI model of erroneous or seemingly fabricated information based on the determination of low confidence scores and limited relevance to the prompt or query.

As noted by researchers at California, Berkeley (Ghose, 2024), "The etiology of AI hallucination includes biased training data, the computational complexity inherent in deep neural networks, lack of contextual / domain understanding, adversarial attack, training on synthetic data ('model collapse'), and a failure to generalize the training data ('overfitting')."

Privacy Issues

Many concerns have arisen due to the geographic distribution of cloud data centers, and users may not necessarily be fully aware or in control of where data is sent or stored. The laws and regulations concerning privacy and compliance may differ between the source of the data, and the cloud data center being used.

Interactions with Other Systems

An obvious concern with interactions by a trained AI system is whether those interactions could impact the behavior of the system. The famous case of an early chatbot named Tay, introduced by Microsoft in 2016, highlighted the potential for interactions with the model after deployment that can re-train the underlying model, thereby affecting its vocabulary and perceived "behavior" (Wolf, et al., 2017).

API Security

Limiting access to an AI system reflects concerns with access, as discussed in the previous vulnerability area. The issue of insecure plugin design and the lack of limiting factors for access control of the application programming interface(s) used by a system and any additional plugins can create an issue where an attacker could achieve unauthorized access.

Posturing

In terms of the applicability of AI in cybersecurity, this contested space can be examined in terms of offensive ("Red Team") and defensive ("Blue Team") techniques. A list of actions applicable to LLMs in each role is shown below.

Red Team Techniques

The potential for the use of AI in adversarial security operations exists in areas like the following:

- Impersonation
- Manipulation
- Disinformation
- Distraction

Indeed, phishing email campaigns readily fit into the first three operations, which became the first application used to highlight security concerns relating to the use of AI.

Additional functions can include the following:

- Creating vulnerable code to create watering hole attacks
- Maintaining and directing botnets

- System “fuzzing”
- Mimicking voices
- Malware development

Blue Team Techniques

These techniques can include the following:

- Using predictive analytics
- Identifying insecure coding patterns
- Eliminating backdoors
- Detecting unauthorized access to a system
- Malware identification
- Data loss prevention

Application Areas

Applications for the use of AI in cybersecurity include access log analysis, phishing prevention, malspam email detection, and the automation of malware analysis.

Conventional attacks on software supply chains, such poisoning of models and data, waterhole attacks, and virtual “dumpster diving” have been applied to AI development efforts. Some examples include code repository scanning for secrets stored in code (e.g., on TorchHub, HuggingFace, and GitHub), including passwords, API keys, and access tokens.

Researchers at Cornell, the Israel Institute of Technology, and Intuit posed the research question: “Can attackers develop malware to exploit the GenAI component of an agent and launch a cyber-attack on the entire GenAI ecosystem?” (Cohen, et al., 2024) They proved that the generative AI models Gemini Pro, ChatGPT 4.0, and LLaVA were indeed susceptible to spamming and exfiltration of personal data from AI-enabled email assistants in several use cases through the use of the worm, named Morris II, that they developed.

There is also concern about the use of AI to enable automated adversarial attacks on critical infrastructure, such as the industrial control systems (ICS) used in communications, finance, and the processing of water, wastewater, energy, and other essential functions, as discussed by asset owners and operators, professional organizations, and civil and military agencies.

Chapter 7: Conclusion

While aspects of AI and ML have been researched and used for decades, recent asymptotic growth of the technologies, especially in areas such as Generative AI and pattern recognition, has led to a new wave of technologies that could significantly reshape real-time system operations in electric power. These technologies are appearing at a critical moment in which the BPS is subject to many dynamic changes, including the growth of variable and distributed energy resources, growing challenges due to changing demand and weather, and increasing regulatory concerns.

Because the BPS is both one of the most critical and complex infrastructures on the planet, ensuring that these technologies are implemented correctly, and function well, is critically important. Beyond simply a technology problem, this is more importantly about building a strong bridge between the humans tasked with the grid's reliability and resilience and these AI/ML technologies to allow the creation of a new, more advanced team that synergistically builds on humans' and AI's distinct strengths.

The electric power sector has no tolerance for significant "trial and error" learning and needs to avoid the "initial bumpy road" observed when new technologies are brought into real time, such as the increase in fatal airline crashes following the introduction of autopilot and the various highly impactful yet incorrect recommendations made by predictive systems in recent years. This is not to say that any system, whether technological or human, can ever be truly perfect or error free. Rather, lessons learned from these previous implementations should be leveraged to ensure a smoother implementation.

The future of the BPS is bright, and the collaboration of humans and machines allows us to move into it confidently.

Appendix A: Bibliography

Further Reading and Recommended Resources

Recent months have shown a significant leap forward in several aspects of AI/ML technologies, in particular generative and predictive technologies. Because of the recognition of technology's importance and relevance to real-time operations, several recent works are referenced within this work. This work does not seek to replicate or repeat those aspects but rather to build upon them as they relate to successful implementations in real-time operations.

While many public works and news sources are cited in this document, the following are of particular relevance to the issues of real-time operations' use of AI/ML and their integration in organizations:

- Department of Energy (2024, April). AI for Energy: Opportunities for a Modern Grid and Clean Energy Economy.
  - https://www.energy.gov/sites/default/files/2024-04/AI%20EO%20Report%20Section%205.2g%28i%29_043024.pdf
- Department of Energy / CESER (2024, April). Potential Benefits and Risks of Artificial Intelligence for Critical Energy Infrastructure.
  - https://www.energy.gov/sites/default/files/2024-04/DOE%20CESER_EO14110-AI%20Report%20Summary_4-26-24.pdf
- Department of Energy (Draft). Human and Organizational Performance Improvement Handbook, Volume 2.
  - (Prior version: https://www.standards.doe.gov/files/doe-hdbk-1028-2009-human-performance-improvement-handbook-volume-2-human-performance-tools-for-individuals-work-teams-and-management)
- Idaho National Labs (2017). Cyber-Informed Engineering
  - https://inldigitallibrary.inl.gov/sites/sti/sti/7323660.pdf
- IEEE (2016). Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems.
  - https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf
- National Cyber Security Centre (2024, April). Deploying AI Systems Securely.
  - https://media.defense.gov/2024/Apr/15/2003439257/-1/-1/0/CSI-DEPLOYING-AI-SYSTEMS-SECURELY.PDF
- National Institute of Standards and Technology (2024, April). Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: An SSDF Community Profile (Initial Public Draft).
  - https://doi.org/10.6028/NIST.SP.800-218A.ipd
- Pacific Northwest National Labs (2024, March). Artificial Intelligence/Machine Learning Technology in Power System Applications.
  - https://www.pnnl.gov/main/publications/external/technical_reports/PNNL-35735.pdf
- White House (2023, October). Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence.
  - https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

References

Abrams, Z. (2024). Addressing equity and ethics in artificial intelligence. Monitor on Psychology, 55(3), 24–34

Anderson, R. S., Benjamin, J., Wright, V. L., Quinones, L., & Paz, J. (2017). Cyber-informed engineering (No. INL/EXT-16-40099). Idaho National Laboratory (INL), Idaho Falls, ID (United States).

Brewster, T. (2024, March). Hackers breached hundreds of companies' AI servers, researchers say. Forbes. https://www.forbes.com/sites/thomasbrewster/2024/03/26/hackers-breach-hundreds-of-ai-compute-servers-researchers-say/

Chakraborti, T., Kambhampati, S., Scheutz, M., & Zhang, Y. (2017). AI challenges in human-robot cognitive teaming. arXiv preprint arXiv:1707.04775.

Cloud Security Alliance (2024, June). Large Language Model (LLM) Threats Taxonomy. https://cloudsecurityalliance.org/artifacts/csa-large-language-model-llm-threats-taxonomy

Cochran, A., & Rayo, M. F. (2023, March). Toward joint activity design: augmenting user-centered design with heuristics for supporting joint activity. In Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care (Vol. 12, No. 1, pp. 19–23). Sage CA: Los Angeles, CA: SAGE Publications

Cohen, S., Bitton, R., & Nassi, B. (2024). Here Comes the AI Worm: Unleashing Zero-click Worms that Target GenAI-Powered Applications. arXiv preprint arXiv:2403.02817.

Cole, D. (2020). The Chinese Room Argument. Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/chinese-room/

Comiter, M. (2019, August). Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It. Harvard University Kennedy School Belfer Center. https://www.belfercenter.org/publication/AttackingAI/

Cummings, M. (2004). Automation Bias in Intelligent Time Critical Decision Support Systems. Collection of Technical Papers - AIAA 1st Intelligent Systems Technical Conference. 2. 10.2514/6.2004-6313.

Daniels, O.J., Murdick, D. (2024, July). Enabling Principles for AI Governance. Georgetown Center for Security and Emerging Technology. https://cset.georgetown.edu/publication/enabling-principles-for-ai-governance/

Deloitte (2024, August). Deloitte's State of Generative AI in the Enterprise Quarter three report. https://www2.deloitte.com/content/dam/Deloitte/us/Documents/consulting/us-state-of-gen-ai-q3.pdf

Department of Energy (2024, April). AI for Energy: Opportunities for a Modern Grid and Clean Energy Economy. https://www.energy.gov/sites/default/files/2024-04/AI%20EO%20Report%20Section%205.2g%28i%29_043024.pdf

Department of Energy / CESER (2024, April). Potential Benefits and Risks of Artificial Intelligence for Critical Energy Infrastructure. https://www.energy.gov/sites/default/files/2024-04/DOE%20CESER_EO14110-AI%20Report%20Summary_4-26-24.pdf

Dubey, A., Abhinav, K., Jain, S., Arora, V., & Puttaveerana, A. (2020, February). HACO: a framework for developing human-AI teaming. In *Proceedings of the 13th Innovations in Software Engineering Conference on Formerly known as India Software Engineering Conference* (pp. 1–9).

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. Human factors, 37(2), 381–394.

European Union (2024, July). REGULATION (EU) 2024/1689 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 13 June 2024 (Artificial Intelligence Act). THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION. Official Journal of the European Union. Amended July 12, 2024.
https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

Flynn, J. R. (2013). Intelligence and human progress: The story of what was hidden in our genes. London: Elsevier.

Garcia, B.F. (2023, November). Ray, versions 2.6.3, 2.8.0. BishopFox. https://bishopfox.com/blog/ray-versions-2-6-3-2-8-0

Ghose, S. (2024, May). Why hallucinations matter: misinformation, brand safety and cybersecurity in the age of Generative AI. University of California Berkeley College of Engineering ACET. https://scet.berkeley.edu/why-hallucinations-matter-misinformation-brand-safety-and-cybersecurity-in-the-age-ofgenerative-ai/

Goldstein, J.A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., Sedova, K. (2023, Jan). Generative language models and automated influence operations: Emerging threats and potential mitigations. https://cdn.openai.com/papers/forecasting-misuse.pdf

Halverson, S. & Szczepanski, N. (2019, Jan). Human Performance in the Control Room Task Force (HPCRTF) – 2018 Distractions in the Control Room Project. Western Electricity Coordinating Council.

Holdsworth, J. & Scapicchio, M. (2024, June). What is deep learning? https://www.ibm.com/topics/deep-learning

Hollnagel, E., & Woods, D. D. (2005). Joint cognitive systems: Foundations of cognitive systems engineering. CRC press.

Hollnagel, E., Woods, D. D., & Leveson, N. (Eds.). (2006). Resilience engineering: Concepts and precepts. Ashgate Publishing Ltd.

Hollnagel, E., & Woods, D. D. (1983). Cognitive systems engineering: New wine in new bottles. International journal of man-machine studies, 18(6), 583–600.

Hollnagel, E. (2009). The four cornerstones of resilience engineering. In Resilience Engineering Perspectives, Volume 2 (pp. 139–156). CRC Press.

IEEE. (2019). Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems (1st ed.). IEEE Standards Association. Retrieved from https://standards.ieee.org/wp-content/uploads/import/documents/other/ead_v2.pdf

Jones, H. (2021, July). Going beyond reliability to robustness and resilience in space life support systems. 50th International Conference on Environmental Systems.

Jones, M.T. (2017, December). Models for Machine Learning. https://developer.ibm.com/articles/cc-models-machine-learning/

Jorge, C. C., Jonker, C. M., & Tielman, M. L. (2023). Artificial trust for decision-making in human-AI teamwork: Steps and challenges. In Proceedings of the HHAI-WS 2023: Workshops at the Second International Conference on Hybrid Human-Artificial Intelligence (HHAI).

Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.

Kerr, K., Norris, T., Stockdale, R. (2007). Data quality information and decision making: a healthcare case study. ACIS 2007 Proceedings, 98.

Kim, Y. H., & Yang, J. J. (2004). What makes circuit breakers attractive to financial markets? A survey. Financial Markets, Institutions & Instruments, 13(3), 109–146.

Klein, G. (2008). Naturalistic decision making. Human factors, 50(3), 456–460.

Klein, G., Feltovich, P. J., Bradshaw, J. M., & Woods, D. D. (2004, June). Common ground and coordination in joint activity. Organizational simulation, 53, 139–184.

Kramer, Mark A. (1991). "Nonlinear principal component analysis using autoassociative neural networks". AIChE Journal. 37 (2): 233–243

Legatt, M. (2015, March). Blackout fried chicken. Paper presented at the NERC Improving Human Performance on the Grid conference, Atlanta, GA

Lopez, J., Textor, C., Lancaster, C., Schelble, B., Freeman, G., Zhang, R., McNeese, N.J. & Pak, R. (2023). The complex relationship of AI ethics and trust in human–AI teaming: insights from advanced real-world subject matter experts. AI and Ethics, 1–21.

Ma, S., Lei, Y., Wang, X., Zheng, C., Shi, C., Yin, M., & Ma, X. (2023, April). Who should I trust: AI or myself? Leveraging human and AI correctness likelihood to promote appropriate trust in AI-assisted decision-making. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (pp. 1–19).

Majumder, S., Dong, L., Doudi, F., Cai, Y., Tian, C., Kalathil, D., Ding, K., Thatte, A., Li, N. & Xie, L. (2024). Exploring the capabilities and limitations of large language models in the electric energy sector. *Joule*, *8*(6), 1544–1549.

McNeese, N. J., Schelble, B. G., Canonico, L. B., & Demir, M. (2021). Who/what is my teammate? Team composition considerations in human–AI teaming. *IEEE Transactions on Human-Machine Systems*, *51*(4), 288–299.

Merlo, J. (2012, October). Human Performance: The Science behind the Tools. https://www.spp.org/documents/18900/humanperformancevideo.pdf

Midcontinent Independent System Operator (2023, Oct). BPM 025 – Operational Forecasting. https://cdn.misoenergy.org/BPM-025%20Operational%20Forecasting49602.zip

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review, 63(2), 81–97. https://doi.org/10.1037/h0043158

Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. Artificial Intelligence Review, 1–66.

MITRE Corporation (2024, March). MITRE ATLAS Framework. ATLAS 4.5.2. https://atlas.mitre.org/matrices/ATLAS

MITRE Corporation (2024, April). MITRE ATT&CK Cloud Framework. ATT&CK v15.1. https://attack.mitre.org/matrices/enterprise/cloud/

Mosier, K. L., & Skitka, L. J. (1999, September). Automation use and automation bias. In Proceedings of the human factors and ergonomics society annual meeting (Vol. 43, No. 3, pp. 344–348). Sage CA: Los Angeles, CA: SAGE Publications.

National Cyber Security Centre (2024, April). Deploying AI Systems Securely. https://media.defense.gov/2024/Apr/15/2003439257/-1/-1/0/CSI-DEPLOYING-AI-SYSTEMS-SECURELY.PDF

National Institute of Standards and Technology (2023, January). NIST Risk Management Framework Aims to Improve Trustworthiness of Artificial Intelligence. https://www.nist.gov/news-events/news/2023/01/nist-risk-management-framework-aims-improve-trustworthiness-artificial

National Institute of Standards and Technology (2023, January). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

National Institute of Standards and Technology (2024, April). Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: An SSDF Community Profile (Initial Public Draft). https://doi.org/10.6028/NIST.SP.800-218A.ipd

NERC (2013, August). Reliability Terminology. https://www.nerc.com/aboutnerc/documents/terms%20aug13.pdf

OpenAI (2023, March). GPT-4 System Card. https://cdn.openai.com/papers/gpt-4-system-card.pdf

OWASP (2024, April). OWASP Top 10 for Large Language Model Applications. Version 1.1. https://owasp.org/www-project-top-10-for-large-language-model-applications/llm-top-10-governance-doc/LLM_AI_Security_and_Governance_Checklist-v1.1.pdf

Pacific Northwest National Labs (2024, March). Artificial Intelligence/Machine Learning Technology in Power System Applications. https://www.pnnl.gov/main/publications/external/technical_reports/PNNL-35735.pdf

Pflanzer, M., Traylor, Z., Lyons, J. B., Dubljević, V., & Nam, C. S. (2023). Ethics in human–AI teaming: principles and perspectives. AI and Ethics, 3(3), 917–935.

Rayo Michael F. (2017). “Designing for collaborative autonomy: updating user-centered design heuristics and evaluation methods.” Proceedings of the Human Factors and Ergonomics Society Annual Meeting. Vol. 61. No. 1. Sage CA: Los Angeles, CA: SAGE Publications

Sawant, S., Mallick, R., McNeese, N., & Chalil Madathil, K. (2022, September). Mutually beneficial decision making in Human-AI teams: Understanding soldier’s perception and expectations from AI teammates in human-AI teams. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 66, No. 1, pp. 287–289). Sage CA: Los Angeles, CA: SAGE Publications.

Schelble, B. G., Lancaster, C., Duan, W., Mallick, R., McNeese, N. J., & Lopez, J. (2023, January). The Effect of AI Teammate Ethicality on Trust Outcomes and Individual Performance in Human-AI Teams. In HICSS (pp. 322–331).

Schelble, B. G., Lopez, J., Textor, C., Zhang, R., McNeese, N. J., Pak, R., & Freeman, G. (2024). Towards ethical AI: Empirically investigating dimensions of AI ethics, trust repair, and performance in human-AI teaming. Human Factors, 66(4), 1037–1055.

Schelble, Beau & Flathmann, Christopher & Musick, Geoff & McNeese, Nathan & Freeman, Guo. (2022). I See You: Examining the Role of Spatial Information in Human-Agent Teams. Proceedings of the ACM on Human-Computer Interaction. 6. 1–27. 10.1145/3555099.

Schiller, C. A. (2024). The Human Factor in Detecting Errors of Large Language Models: A Systematic Literature Review and Future Research Directions. arXiv preprint arXiv:2403.09743.

Schwartz, O. (2024, January). In 2016, Microsoft’s Racist Chatbot Revealed the Dangers of Online Conversation The bot learned language from people on Twitter—but it also learned values. IEEE Spectrum. https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation

Searle, J. R. (1980). Minds, brains, and programs. Behavioral and brain sciences, 3(3), 417–424.

Shostack, A. (2014). Threat Modeling: Designing for Security. Wiley. Indianapolis, IN. February 17, 2014. ISBN 978-1-118-80999-0.

Smith, P. J., McCoy, C. E., & Layton, C. (1997). Brittleness in the design of cooperative problem-solving systems: The effects on user performance. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 27(3), 360–371.

Snowden, D. J., & Boone, M. E. (2007). A leader's framework for decision making. Harvard business review, 85(11), 68.

Souppaya, M., Scarfone, K., & Dodson, D. (2022). Secure software development framework (ssdf) version 1.1. NIST Special Publication, 800, 218.

Sullenberger, S. (2015, April). Technology cannot replace pilots. https://www.linkedin.com/pulse/technology-cannot-replace-pilots-capt-sully-sullenberger/

Tamari, S., Tzadik, S. (2024, April). Wiz Research finds architecture risks that may compromise AI-as-a-Service providers and consequently risk customer data; works with Hugging Face on mitigations. Wiz Research. https://www.wiz.io/blog/wiz-and-hugging-face-address-risks-to-ai-infrastructure

Xie, L., Zheng, X., Sun, Y., Huang, T., & Bruton, T. (2022). Massively digitized power grid: opportunities and challenges of use-inspired AI. Proceedings of the IEEE, 111(7), 762–787.

Zhang, R., McNeese, N. J., Freeman, G., & Musick, G. (2021). " An ideal human" expectations of AI teammates in human-AI teaming. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW3), 1–25.

Zhang, Rui & Duan, Wen & Flathmann, Christopher & McNeese, Nathan & Freeman, Guo & Williams, Alyssa. (2023). Investigating AI Teammate Communication Strategies and Their Impact in Human-AI Teams for Effective Teamwork. Proceedings of the ACM on Human-Computer Interaction. 7. 1-31. 10.1145/3610072.

Zhou, J., & Chen, F. (2019, August). Towards trustworthy human-AI teaming under uncertainty. In *IJCAI 2019 workshop on explainable AI (XAI)*.

White House (2013). Presidential Policy Directive -- Critical Infrastructure Security and Resilience. https://obamawhitehouse.archives.gov/the-press-office/2013/02/12/presidential-policy-directive-critical-infrastructure-security-and-resil

White House (2023, October). Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/

Woody, C. (2022, March). Incorporating Supply Chain Risk and DevSecOps into a Cybersecurity Strategy. Software Engineering Institute. https://insights.sei.cmu.edu/library/incorporating-supply-chain-risk-and-devsecops-into-a-cybersecurity-strategy/

Wolf, M. J., Miller, K., & Grodzinsky, F. S. (2017). Why we should have seen that coming: comments on Microsoft's tay" experiment," and wider implications. Acm Sigcas Computers and Society, 47(3), 54–64.

Zissis, G. (2019). The R3 Concept: Reliability, Robustness, and Resilience [President's Message]. *IEEE Industry Applications Magazine*, *25*(4), 5–6.

Appendix B: Acronyms

Acronyms Used in Report	
Acronym	**Definition**
AI	Artificial Intelligence
API	Application Programming Interface
ARIMA	Autoregressive Integrated Moving Average
BA	Balancing Authority
BPS	Bulk Power System
CEII	Critical Energy Infrastructure Information
CESER	(DOE) Cybersecurity, Energy Security, and Emergency Response
CNN	Convolutional Neural Networks
CSET	Center for Security and Emerging Technology
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DL	Deep Learning
EDA	Exploratory Data Analysis
EMS	Energy Management System
EMSWG	EMS Working Group
ERO	Electric Reliability Organization
GAN	Generative Adversarial Network
GPT	Generative Pre-Trained Transformer
GRU	Gated Recurrent Units
HOP	Human and Organizational Performance
HPI	Human Performance Improvement
ICS	Industrial Control Systems
JIT	Just In Time
KNN	K-Nearest Neighbors
KPI	Key Performance Indicators
LBA	Local Balancing Authority
LLM	Large Language Model
LSTM	Long Short-Term Memory

ML	Machine Learning
MTLF	Mid-Term Load Forecast
NERC	North American Electric Reliability Corporation
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
NOC	Network Operations Center
OT	Operational Technology
RAG	Retrieval Augmented Generation
RC	Reliability Coordinator
RE	Regional Entity
RNN	Recurrent Neural Networks
RTOS	Real-Time Operating Subcommittee
RTU	Remote Terminal Unit
SCADA	Supervisory Control and Data Acquisition
SE	State Estimation
SOC	Security Operations Center
STLF	Short-Term Load Forecast
SVD	Singular Value Decomposition
SVM	Support Vector Machine
SVR	Support Vector Regression
VAE	Variational Autoencoder

Appendix C: Contributors

Contributors to this whitepaper include members of NERC, E-ISAC, the ERO Enterprise, SITES, SWG, EMSWG, and other industry stakeholders. Contributions include research, discussion, writing, and editing. A special thank you goes to ResilientGrid staff for their significant collaborative efforts and contributions to this whitepaper. In alphabetical order, the list of contributors include the following individuals:

Whitepaper Contributors		
Name	**Title**	**Organization**
Christopher Pilong	Senior Director, Operations Planning	PJM Interconnection, L.L.C.
Brandon Turner	Bulk Power System Awareness Analyst	North American Electric Reliability Corporation
Darrell Moore	Director, Bulk Power System Awareness and Personnel Certification	North American Electric Reliability Corporation
Dwayne Fewless	Principal Analyst	ReliabilityFirst
James (Jimmy) Hartmann	Senior Director, Control Room Operations and Operations Training	Electric Reliability Council of Texas, Inc.
Joseph Januszewski	Senior Cyber Security Analyst	North American Electric Reliability Corporation – E-ISAC
Lance Ransom	COO and Co-Founder	ResilientGrid, Inc.
Michael Legatt	CEO and Co-Founder	ResilientGrid, Inc.
Rob Adams	Executive Director, Power Delivery - Smart Grid & Innovation	Florida Power & Light
Stephanie Lawrence	Senior Program Specialist	North American Electric Reliability Corporation
Syedkhair Quadri	Advisor, R&D and Strategic Ventures	Midcontinent ISO
Timothy Beach	Director Reliability Coordination	California Independent System Operator Corporation
Wei Qiu	Lead Engineer of Event Analysis	North American Electric Reliability Corporation