DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. Reference herein to any social initiative (including but not limited to Diversity, Equity, and Inclusion (DEI); Community Benefits Plans (CBP); Justice 40; etc.) is made by the Author independent of any current requirement by the United States Government and does not constitute or imply endorsement, recommendation, or support by the United States Government or any agency thereof.

SANDIA REPORT
SAND2025-00392
Unclassified Unlimited Release
Printed January 2025

Sandia National Laboratories

Application of Artificial Intelligence/Machine Learning to Operations Research

Taylor McKenzie, Kelsey Abel, John Flory, Angie Kelic, Marilee Orr, Trey Reilly

Controlled by: Sandia National Laboratories

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico
87185 and Livermore,
California 94550

Issued by Sandia National Laboratories, operated for the United States Department of Energy by National Technology & Engineering Solutions of Sandia, LLC.

ABSTRACT

This report examines the transformative impact of Artificial Intelligence (AI) and Machine Learning (ML) on operations research, private industry, and government sectors, highlighting their applications in automating processes, enhancing decision-making, and optimizing complex systems. AI/ML technologies have revolutionized industries through predictive maintenance, supply chain optimization, and autonomous systems, while also advancing public safety and defense operations. However, challenges such as data integrity, model transparency, and the need for human oversight persist, particularly in high-consequence environments. The report emphasizes the critical role of explainable AI (XAI) and human-computer interaction models like Human-in-the-Loop (HITL) and Human-on-the-Loop (HOTL) in fostering trust and accountability. Balancing automation with ethical responsibility and transparency is essential for the continued successful integration of AI/ML into operational and strategic decision-making frameworks.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

This page left blank

EXECUTIVE SUMMARY

Artificial Intelligence (AI) and Machine Learning (ML) have become transformative forces across private industry and government sectors, reshaping fields like operations research, logistics, and decision-making. In private industry, AI/ML applications have revolutionized operations management, automating processes such as scheduling, predictive maintenance, and supply chain optimization. Operations research, which traditionally focuses on decision support and optimization, has increasingly integrated AI/ML techniques to enhance the scale and accuracy of complex problem-solving tasks. Government sectors, particularly in defense and public safety, have similarly employed AI/ML to enable real-time decision-making in high-consequence environments, although these applications often require a balance between automation and human oversight.

Despite the advantages of AI/ML in operations research and other fields, several challenges persist. Ensuring data integrity, such as addressing missing or incomplete data, remains critical to maintaining the reliability of AI/ML models. This is especially important in high-stakes environments like autonomous vehicles and air traffic control, where even small errors can have severe consequences. Another significant challenge is the transparency and interpretability of AI/ML models, often referred to as "black box" systems. While operations research models have traditionally emphasized explainability, many AI/ML systems lack this clarity, making it difficult for human operators to fully trust or understand AI-driven decisions. Efforts in explainable AI (XAI) are progressing but further development is necessary to ensure that AI models can provide interpretable and actionable insights.

As AI/ML continues to evolve, its integration with operations research and other domains will depend on maintaining effective human-computer interaction. Human-in-the-loop (HITL) and human-on-the-loop (HOTL) systems are essential to preserving human oversight while leveraging AI's ability to process vast amounts of data and optimize decisions. For long-term success, AI/ML systems must strike a balance between automation and human involvement, ensuring transparency, reliability, and ethical responsibility in both high-consequence environments and broader industry applications.

ACRONYMS AND TERMS

Acronym/Term	Definition
AI	Artificial Intelligence
ANOVA	Analysis of Variance
DSS	Decision Support Systems
EU	European Union
GBT	Gradient Boosted Trees
HITL	Human in the Loop
HOTL	Human on the Loop
ML	Machine Learning
NN	Neural Network
OECD	Organization for Economic Cooperation and Development
OM	Operations Management
OR	Operations Research
RF	Random Forest
SVM	Support Vector Machine
UAV	Unmanned Aerial Vehicle
VAR	Vector Autoregressive
XAI	Explainable Artificial Intelligence

1. INTRODUCTION

Artificial Intelligence (AI) and Machine Learning (ML) have become pivotal technologies in the 21st century, reshaping industries by automating complex tasks, delivering actionable insights from large datasets, and enabling more efficient decision-making processes that were once exclusively human. The increasing availability of data, alongside advancements in computational power and algorithmic development, has allowed AI/ML to transcend academic research and become cornerstones of modern industry and government functions.

AI and ML are distinct yet interconnected fields, each with its own evolution and methodologies. Traditional statistical modeling laid the groundwork for AI/ML development, with an emphasis on descriptive and predictive modeling. AI, broadly defined as systems that can make predictions or decisions within human-defined parameters, often incorporates machine learning as a core mechanism. Machine learning models are designed to learn from data, adapting their decision-making capabilities without being explicitly programmed for every scenario. This flexibility makes ML particularly valuable in environments that involve large volumes of data or where patterns are difficult for humans to discern unaided.

Within private industry, AI/ML applications have revolutionized fields such as operations management, logistics, manufacturing, and finance. These technologies have enabled companies to automate processes that humans previously conducted, such as scheduling, inventory management, and quality control, increasing both speed and scale. AI has also been used in predictive maintenance, allowing industries to reduce downtime by forecasting when machines are likely to fail. More advanced applications, such as human-robot collaboration in manufacturing and reinforcement learning in supply chain optimization, continue to push the boundaries of what AI/ML systems can achieve.

Government sectors are also leveraging AI/ML for various applications, particularly in areas like defense, public safety, and transportation. High-consequence systems, such as autonomous vehicles and air traffic control, rely on AI/ML for real-time decision-making. However, the integration of these systems comes with significant challenges, such as ensuring accuracy, maintaining human oversight, and safeguarding critical infrastructure. The human-computer interaction in these high-stakes environments highlights the need for explainable AI, where humans must understand and trust the decisions being made by machines.

Despite the advantages AI/ML brings to both industry and government, challenges remain. One critical concern is the management of missing or incomplete data, which can compromise the accuracy and reliability of AI/ML models. Additionally, while AI systems have proven capable of replicating many human tasks, certain high-stakes environments still require human intervention, either to provide oversight or to handle novel situations that AI/ML systems cannot generalize from past experiences. This human-in-the-loop model is essential in areas where errors can have catastrophic consequences, such as autonomous vehicles or defense systems.

The purpose of this report is to explore the current and emerging applications of AI and ML across private industry and government. We will review both the benefits and limitations of these technologies, with particular attention paid to the challenges of data integrity, explainability, and human-computer teaming. By analyzing these applications, we aim to provide a comprehensive overview of how AI/ML systems are transforming operational efficiency and decision-making processes, while also considering the ethical and technical challenges they introduce.

2. ARTIFICIAL INTELLIGENCE/MACHINE LEARNING FUNDAMENTALS

The fields of artificial intelligence (AI) and machine learning (ML) have seen rapid growth in recent decades, reshaping the landscape of data-driven decision-making across numerous sectors. Despite their interrelated nature, AI and ML draw upon different methodologies and approaches, often in conjunction with traditional statistical models. Understanding the distinctions and overlaps between these fields is crucial for determining how they can be effectively applied to solve specific problems. This section provides an overview of the evolution of AI/ML, discusses their relationship to statistics, and outlines their respective roles in modeling, prediction, and decision-making.

2.1. Evolution of the Field and Definitions

Statistics, machine learning, and artificial intelligence bear many similarities, including foundational theories they draw on, methods they employ, and problems they are designed to address. Each of these fields, however, has also seen its own independent developments geared towards specific and specialized techniques and applications. As a result, it can be difficult to delineate these disciplines and identify what methods are most appropriate for a given problem. This section compares and contrasts views of statistics, ML, and AI commonly seen in existing literature and considers strengths and weaknesses of each discipline.

Broadly, statistical modeling can be grouped into two "cultures" (Breiman, 2001). The first culture is concerned with descriptive and/or explanatory modeling, providing insights into how events played out, the current state of the environment, and key relationships observed in historic data (Boulesteix & Schmid, 2014). This approach assumes that the data generating process, a unified description of relationships between inputs and outputs and of uncertainty in data, is known to the practitioner, though parameters of the data generating process may be unknown. Through the use of traditional statistical methods like descriptive statistics, linear regression, and ANOVA,[1] practitioners can estimate relationships between inputs and outputs and precisely describe uncertainty in those estimates. Experimental data are clearly useful to this culture of modeling, since the conditions by which data are generated and observed can be understood and carefully controlled for. Other disciplines in this culture, like economics and the social sciences, leverage "quasi-experimental" data, which is not produced in an experimental setting but has characteristics that allow it to be used as experimental data, often following some statistical corrections. While traditional statistics models leveraged by this culture are almost exclusively estimated using computational resources and procedures, these methods are usually not considered to be "machine learning".

The second culture of statistical modeling is concerned with making predictions rather than describing or explaining phenomena. As a result, highly-structured models that include detailed information about the data generating process are unnecessary. Instead, this culture frequently employs highly-flexible models that very closely approximate the data they were trained on. To ensure models are not overfitted (i.e., memorizing the data rather than learning the relationships), split sample techniques like cross-validation are frequently used. In this approach, the dataset is divided into two subsets: one for training the model and the other for testing its performance, providing an unbiased estimate of the model's predictive accuracy on unseen data. Training these models with non-experimental data may result in models that do not capture causal relationships, which could introduce bias into predictions. Nonetheless, these approaches are frequently used on

[1] ANOVA, or ANalysis Of VAriance, is a statistical method used to determine whether there are significant differences between the means of three or more independent groups.

massive non-experimental datasets and produce predictions that have operational value. Since these approaches do not have a predefined structure and typically include all variables that could potentially improve predictions, techniques to reduce the dimensionality of inputs to the model are sometimes used. Table 1 lists analytical problems and associated modeling approaches that commonly appeared in research covered in this literature review.

"Artificial intelligence" is a broad term that has been interpreted and applied in many different ways. For example, the Organization for Economic Cooperation and Development (OECD)provides a very broad definition of AI: "A machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments…AI systems are designed to operate with varying levels of autonomy" (OECD, 2016). Generally, ML and statistics models and techniques are the tools with which AI systems automate certain aspects of decision-making that were previously delegated to humans (Molnar, Casalicchio, & Bischl, 2020). As a result, the function of AI for a given decision-making application and associated risks are very similar to delegating aspects of decision-making to human analysts and actors. In particular, humans are granted some (but not full) autonomy in their roles. The level of autonomy granted is ideally selected by superiors to grant analysts some freedom to pursue lines of inquiry and even enact limited policies, but also to ensure that no one individual can cause substantial impacts and that a wider range of stakeholders/decision-makers are consulted for especially consequential decisions. For example, (Brose, 2020) considers autonomous warfighting systems and notes that human soldiers are granted limited and carefully-defined autonomy over their actions, allowing them to make certain decisions and take certain actions under specific circumstances but requiring involvement of commanding officers in other situations. The author makes the case that AI warfighting systems must be governed similarly, granting a limited level of autonomy to perform defined actions in certain situations but requiring outside intervention otherwise. Of course, AI systems process information differently than humans, raising concerns about explainability and interpretability. These characteristics of AI systems and how they compare and interact with human systems is discussed in greater detail in Section 3.3.

Table 1: Common Analytical Problems and Modeling Approaches

Problem	Modeling approach	Example approach(es)
Direct causal inference	Statistics	Regression (e.g., linear, logistic)
Indirect causal inference (e.g., using quasi-experimental data)	Statistics	Regression (e.g., two-stage least squares)
Estimating impacts of specific policies (retrospective and/or prospective)	Statistics	Regression (e.g., difference-in-difference), regression discontinuity, time-series modeling (e.g., vector autoregression, structural models)
Classification into discrete categories	ML	Support vector machines, k-means clustering, random forests, neural networks

Problem	Modeling approach	Example approach(es)
Dimensionality reduction	ML	Principal component analysis, ridge regression
Prediction (e.g., real-valued, categorical)	Statistics or ML	Statistics: Regression (e.g., linear, logistic, multinomial), structural models ML: Neural networks, Gaussian processes, support vector machines

2.2. Applications of AI to Operations Research and Management

OR is a field of study that emerged in World War II and has seen tremendous growth through the information age. Broadly, OR aims to support decision-making through robust data-driven analyses, striving to be a "scientific method of providing executive departments with a quantitative basis for decisions regarding the operations under their control" (Morse & Kimball, 1947). OR analyses can be either descriptive (to better understand functions and dynamics of systems and processes), predictive (predicting and estimating the future state of systems and processes given current states), or normative (prescribing a course of action based on current states, objectives, and constraints) (Wacker, 1998). OR analyses leverage decision support systems (DSSs), which includes information, computation, and knowledge infrastructure, to translate data, develop insights, and ultimately inform actions (Sprague Jr., 1980). Generally, DSSs tend to be data-focused (relying on observational data frequently stored in databases), knowledge-focused (relying on background and context to establish system rules), or model-based (relying on mathematical and computational models) (Holsapple, 2008).

AI and ML techniques have offered enormous potential for improving efficiency and effectiveness of operations management (OM). In the most successful cases, AI/ML has been leveraged to (1) automate processes that humans formerly conducted and perform those functions more quickly and at larger scales, (2) process large amounts of varied data to produce insights that were difficult for humans to glean, (3) enable machines/robots that operate more efficiently/effectively than humans and/or in environments inhospitable to humans. However, AI/ML approaches have not been able to effectively replace critical components of human decision-making. This section reviews literature on AI/ML applications in operations research (OR) and operations management (OM), with a focus on enumerating tasks for which machines could provide substantial benefits vs. tasks for which human involvement has continued to be necessary.

There is significant overlap between the capabilities of AI/ML and the needs of OR and DSSs. (Gupta, Modgil, Bhattacharyya, & Bose, 2022) developed a taxonomy to classify AI approaches, DSS focus, and OR analysis type, shown in Figure 1. The authors note a few general areas in which AI has successfully contributed to DSSs and OR: (1) Development and deployment of expert systems that synthesize knowledge bases and make information easily retrievable to researchers and decision-makers, (2) ML techniques, especially those that can process large amounts of disparate data to produce insights, and (3) Natural language processing and related techniques that are capable of analyzing unstructured data. Each of these strengths is an area in which AI/ML can effectively replicate human activities but can do so faster, at larger scales, and/or with lower error rates.

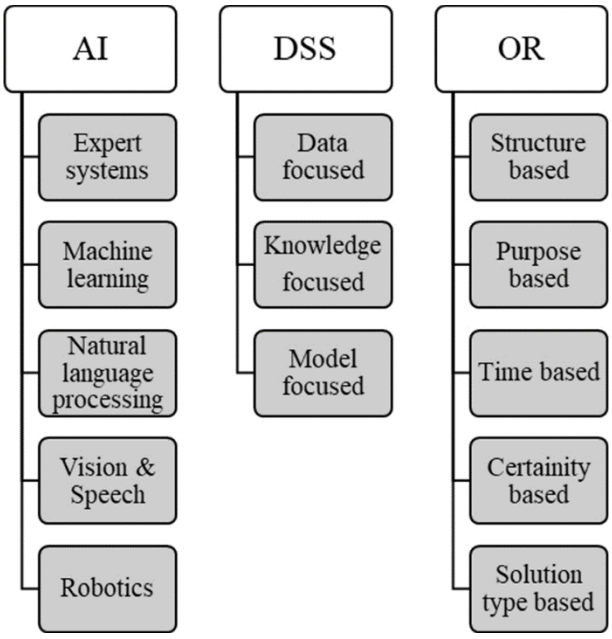

Figure 1: AI, DSS, and OR Taxonomy (Gupta, Modgil, Bhattacharyya, & Bose, 2022)

(Fahle, Prinz, & Kuhlenkotter, 2020) reviewed a litany of ML methods and assessed their applicability to a number of modern manufacturing processes. The authors' high-level summary is replicated in Table 2. This review reveals that ML methods can be brought to many aspects of manufacturing and OM in general. In particular, neural networks appear to be especially useful for predictions of complex system outcomes like costs and production schedules, while clustering and classification methods and decision trees appear especially useful for production and logistics management. Further, most ML methods commonly used in current manufacturing processes are supervised methods, meaning that humans must classify the data that ML algorithms are trained on. However, reinforcement learning algorithms, which take a trial-and-error approach similar to human learning, have seen growing interest in recent years.

Table 2: Manufacturing Applications and Algorithms (replicated from (Fahle, Prinz, & Kuhlenkotter, 2020)

Subtopic	Application	Algorithm(s)
Manufacturing process planning	Scheduling	Q-learning, random forest (RF), decision trees
	Cost and energy prediction	Neural network (NN), support vector machine (SVM), gradient-boosted trees (GBT), RF
	System modeling	Logistic regression, RF, decision trees, Bayesian network
Quality control	Quality cost reduction	Decision trees, NN, SVM
	Process line quality	Decision trees, Bayesian network
Predictive maintenance	Remaining useful life	Decision tree, NN, principal components analysis
Logistics	Scheduling	NN, Q-learning, RF

Subtopic	Application	Algorithm(s)
Robotics	Human-robot collaboration	Hidden Markov model (HMM), K nearest neighbors (KNN), clustering, NN
	Path planning	KNN, NN
Assistance and learning systems	Assembly assistance	NN
AI-training concepts in learning factories	Object recognition	NN
Process control and optimization	Production line	GBT
	Process and tool condition forecast	NN, decision trees, RF, SVM, regression

(Woschank, Rauch, & Zsifkovits, 2020) provide a review of how AI/ML methods could be applied to next generation "smart logistics" and "Industry 4.0", which focus on modernizing industrial manufacturing through increased interconnectivity, digitalization, and automation. The authors found that the plurality of modern research (~42% of published papers) was focused on cyber-physical systems for logistics and predictive maintenance. Other focal areas included improvement of operational logistics and intelligent transport logistics (~24% of published papers) and strategic and tactical process optimization (~12% of published papers). Overall, these reviews indicate that there is significant and ongoing development of methods to bring AI/ML methods to industry and OM of the future.

Several studies have noted the promise of AI/ML approaches for supply chain and logistics management. (Pournader, Ghaderi, Hassanzadegan, & Fahimnia, 2021) reviewed AI applications in supply chain management and identified several clusters of similar approaches based on their bibliometric analysis. The authors grouped approaches into three main AI functions: Decision making (including planning, modeling/simulation, scheduling, and optimization), learning (analyzing frequently large and disparate data to understand behavior and/or make predictions), and hybrid approaches (integrating learning and decision-making functions), listed in Table 3.

Table 3: AI/ML For Supply Chain Management (with info from (Pournader, Ghaderi, Hassanzadegan, & Fahimnia, 2021))

Function	Approach clusters	Example method(s)
Decision making	Simulation and system dynamics	Discrete-event simulation, system dynamics models
	Genetic algorithms and agent-based modeling	Multi-agent/game theory models, mixed integer programming
	Stochastic programming	Robust optimization, regression
Learning	Time-series analysis	Vector autoregressive (VAR) models, spectral analysis
	Big data analytics	Deduction graphs

Function	Approach clusters	Example method(s)
	Neural networks and support vector machines	
Hybrid	AI methods for sustainable SCM	Delphi method, fuzzy cognitive maps
	AI methods for supply chain risk management	

3. IMPLEMENTATION CONSIDERATIONS

In the rapidly evolving field of AI/ML applications, ensuring the integrity and reliability of data is critical to developing robust models. However, real-world data is often incomplete, leading to challenges that can impair model performance. This section addresses these concerns by exploring the different types of missing data, their potential impacts on AI/ML systems, and the strategies available to mitigate these challenges. Additionally, the role of human interactions with AI and the need for explainability in AI/ML models are highlighted, emphasizing the importance of trust and transparency in the successful implementation of AI solutions. These discussions aim to provide a comprehensive understanding of both technical and human-centered considerations essential for advancing AI/ML technologies in complex environments.

3.1. Missing Data in AI/ML Applications

In AI and ML applications, the integrity and completeness of data are paramount to the development of robust and accurate models. However, missing data is an omnipresent challenge that can significantly impair the performance and reliability of these models. Missing data can arise from various sources, including sensor malfunctions, human error, or limitations in data collection processes. Addressing this issue is crucial, as the presence of incomplete data can lead to biased estimates, reduced statistical power, and ultimately, flawed decision-making. This section delves into the types of missing data, their potential impacts on AI/ML applications, and the strategies employed to mitigate these challenges, ensuring the development of more resilient and effective models.

Missing data can be broadly categorized into three classes based on the mechanism of their occurrence: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR). These classifications, as detailed by (Litte & Rubin, 2002) and further elaborated in the seminal work by (Schafer & Graham, 2002), provide a framework for understanding the underlying patterns of missingness and inform the appropriate strategies for handling them.

Missing Completely at Random (MCAR) occurs when the probability of data being missing is independent of both observed and unobserved data. In other words, the missingness is entirely random and does not depend on any variables within the dataset. For instance, if a sensor occasionally fails to record data due to random technical glitches, the resulting missing data can be considered MCAR. When data are MCAR, the analysis remains unbiased, although the statistical power may be reduced due to the smaller sample size (Schafer & Graham, 2002).

When data are Missing Completely at Random (MCAR), the missingness is independent of both observed and unobserved data, making it the simplest type of missing data to handle. One common method to address MCAR is listwise deletion, where any record with missing values is excluded from the analysis. This approach is straightforward and maintains the integrity of the dataset, but it can lead to a significant reduction in sample size, potentially impacting the statistical power of the analysis (Litte & Rubin, 2002). Another method is mean imputation, where missing values are replaced with the mean of the observed values for that variable. While this method is easy to implement, it can underestimate the variability and lead to biased parameter estimates (Schafer & Graham, 2002). More sophisticated techniques, such as multiple imputation, can also be used for MCAR data. Multiple imputation involves creating several complete datasets by imputing missing values multiple times and then combining the results to account for the uncertainty associated with the imputed values (Rubin, 1988).

Missing at Random (MAR) describes a scenario where the probability of missing data is related to observed data but not to the unobserved data. For example, in a medical study, if older patients are less likely to respond to follow-up surveys, the missingness is related to the age variable, which is observed. Under the MAR assumption, the missing data mechanism can be accounted for by conditioning on the observed data (Litte & Rubin, 2002).

For data that are Missing at Random (MAR), the probability of missingness is related to the observed data but not the unobserved data. This allows for more advanced imputation methods that leverage the relationships within the observed data. Multiple imputation is particularly effective for MAR data, as it creates multiple datasets with different imputed values based on the observed data, and then combines the results to produce estimates that reflect the uncertainty of the missing data (Rubin, 1988). Another method is the Expectation-Maximization (EM) algorithm, which iteratively estimates the missing values by maximizing the likelihood function based on the observed data (Dempster, Laird, & Rubin, 1977). Additionally, model-based approaches, such as using maximum likelihood estimation within a structural equation modeling framework, can be employed to handle MAR data by incorporating the missing data mechanism directly into the model (Baraldi & Enders, 2010).

Missing Not at Random (MNAR) occurs when the probability of missing data is related to the unobserved data itself. This type of missingness is particularly challenging because the missingness mechanism is inherently tied to the missing values. For instance, in a survey on income levels, individuals with higher incomes might be less likely to disclose their earnings, leading to a non-random pattern of missing data.

Addressing Missing Not at Random (MNAR) data is more complex, as the missingness is related to the unobserved data itself. One approach to handle MNAR data is to use selection models or pattern-mixture models, which explicitly model the missing data mechanism. Selection models involve specifying a model for the missing data process and then jointly modeling the outcome and the missing data mechanism (Litte & Rubin, 2002). Pattern-mixture models, on the other hand, stratify the data based on the pattern of missingness and then model each stratum separately (Little, 1993). Sensitivity analysis is another crucial technique for MNAR data, where different assumptions about the missing data mechanism are tested to assess the robustness of the results (Molenberghs, Beunckens, Sotto, & Kenward, 2008). In some cases, external data or auxiliary variables that are related to the missingness can be incorporated to inform the imputation process and reduce bias (Collins, Schafer, & Kam, 2001).

In addition to the challenges posed by missing data, AI/ML applications often face the separate but equally critical issue of incomplete utilization of available variables due to data collection or data architecture constraints, such as siloing. Data siloing occurs when data is isolated in separate systems or departments, preventing comprehensive analysis and integration. This fragmentation can lead to suboptimal model performance, as key variables that could enhance predictive accuracy and insights are excluded from the analysis. For instance, in healthcare, patient data might be dispersed across different departments (e.g., radiology, pathology, and primary care), each maintaining its own database with limited interoperability. As a result, crucial variables like imaging results or lab tests may not be incorporated into predictive models, thereby reducing their effectiveness. Addressing these issues requires robust data integration strategies, such as the implementation of data lakes or federated learning approaches, which enable the aggregation and analysis of disparate data sources while maintaining data privacy and security (Kambatla, Kollias, Kumar, & Grama, 2014). By overcoming these architectural barriers, organizations can leverage the full spectrum of available data, leading to more comprehensive and accurate AI/ML models.

By employing these tailored methods to address MCAR, MAR, and MNAR, researchers can mitigate the impact of missing data on their analyses, leading to more accurate and reliable AI/ML models. Understanding the nature of the missing data and selecting appropriate techniques is essential for maintaining the integrity and validity of the results.

3.2. Human/AI Interactions

AI/ML methods have tremendous technical potential, but it is critical that those approaches consider and optimize coordination with human systems to make them most effective. This process includes building and developing with human users in the short term and maintaining trust while staving off complacency in the long term. This section discusses strategies and factors that can improve trust in AI as well as methods to maintain quality human-AI interactions that include human critical thinking.

Previous research has considered factors that make people more or less likely to trust and adopt AI assistance. Multiple studies have conducted experiments looking into the relationship between willingness to accept AI suggestions and participants' confidence in AI and themselves (Schaffer, O'Donovan, Michaelis, Raglin, & Hollerer, 2019). In (Chong, Zhang, Goucher-Lambert, Kotovsky, & Cagan, 2022), study participants were presented with a chess board state and asked to either make their own move or accept a move suggested by an AI player. Participants were also asked to assess confidence in their own chess abilities as well as the AI player's abilities. The study found that individuals' confidence in AI did not significantly affect their propensity to accept or reject AI suggestions. Instead, participants' decision to trust AI was driven by confidence in themselves; in particular, participants who rated their own chess abilities highly tended not to accept AI suggestions, independent of participants' assessment of AI capabilities. Further, the study found that participants who had low confidence in their own abilities tended to attribute poor AI performance and suggestions to themselves, further reducing their self-confidence and making it more likely they will trust AI suggestions (good or bad) in the future. The authors postulate long-term dynamic implications where a dichotomy between those who trust AI suggestions and those who do not is reinforced by interactions with the AI system.

These experiments illustrate the importance of thoughtful integration of AI/ML with human systems. First, as outlined in (Jacovi, Marasovic, Miller, & Goldberg, 2021), it is critical that trust is developed between human users and AI systems. The authors note that in both human-human and human-AI interactions, trust is critically dependent on **anticipation** and **vulnerability**: If Party A anticipates Party B will act in A's interest and if A is willing to accept vulnerability to B's actions, then A trusts B. Thus, in order to build and maintain trust between humans and AI systems, it is necessary that humans can expect AI to behave in predefined ways, and that humans are willing to accept the consequences from any deviation from that expected behavior.

The European Union (EU) outlined several qualities of AI systems that can improve trust in AI systems to both execute their intended tasks and limit negative externalities, listed in Table 4 (EU Commission, 2020). These guidelines cover several technical requirements for trustworthy AI (e.g., technical robustness and safety, transparency, accountability) as well as situations with especially high consequence of undesired behavior (e.g., human agency and oversight, diversity/non-discrimination/fairness, societal and environmental well-being). Notably, explainability and interpretability of AI systems are common themes across factors of trustworthy AI and will be discussed in greater detail in the following section.

Table 4: EU Guidelines for Trustworthy AI

Requirement for Trustworthy AI	Factors
Human agency and oversight	• Foster fundamental human rights • Support users' agency • Enable human oversight
Technical robustness and safety	• Resilience to attack and security • Fallback plan and general safety • A high level of accuracy • Reliability • Reproducibility
Privacy and data governance	• Ensure privacy and data protection • Ensure quality and integrity of data • Establish data access protocols
Transparency	• High-standard documentation • Technical explainability • Adaptable user-centered explainability • Make AI systems identifiable as non-human
Diversity, non-discrimination, and fairness	• Avoid unfair bias • Encourage accessibility and universal design • Solicit regular feedback from stakeholders
Societal and environmental well-being	• Encourage sustainable and eco-friendly AI • Assess the impact on individuals • Assess the impact on society and democracy
Accountability	• Auditability of algorithms/data/design • Minimize and report negative impacts • Acknowledge and evaluate trade-offs • Ensure redress

Longer-term issues that can afflict AI system implementations include complacency and systematic distrust. Without proper controls, individuals can gravitate towards one of these extremes, either placing too much trust in the AI system and taking suggested actions without critical assessment or completely refusing to engage with AI systems (Zerilli, Bhatt, & Weller, 2022). Thoughtful design of AI systems and controls to moderate engagement with these systems maintain "algorithmic vigilance" that lies between complacency and distrust. The ability of AI systems to explain their own reasoning and processing can help build and maintain trust. That said, AI explainability may not reduce complacency as trusting individuals tend to underemploy analytical thinking when presented with AI suggestions and even take explanations (right or wrong) as a signal of AI competence (Bansal, et al., 2021). So-called "cognitive forcing functions" have been proposed as method to increase engagement in human-AI interactions. Specifically, human decision-making relies on a mix of fast, heuristic-based thinking (i.e., System 1 thinking) and slow, deliberate, analytical thinking (i.e., System 2 thinking) (Bucinca, Malaya, & Gajos, 2021). System 1 thinking is typically very fast and can be accurate but can be prone to systematic bias and errors especially as environments or conditions change. By making minor changes to how AI suggestions are presented, it is possible to increase the rate of System 2 thinking. Examples of successful cognitive forcing functions include requiring humans to make a decision prior to seeing AI suggestions (Green & Chen, 2019), delaying the

presentation of AI suggestions (Park, Barber, Kirlik, & Karahalios, 2019), and giving human users the option of whether and when to see AI suggestions (Fitzsimons & Lehmann, 2004).

3.3. AI/ML Explainability

Explainability and interpretability of AI systems serves two purposes. First, as discussed in the previous section, explanations of how the AI system arrived at its findings and recommendations is a major factor in building and maintaining humans' trust in those systems. Second, explainability and interpretability provide a channel through which human users can critically assess AI outputs, though cognitive forcing functions may be necessary to get humans to engage in System 2 thinking. This section discusses methods that have been developed to increase the explainability and interpretability of AI systems with the goal of improving human-AI interactions and teaming.

It is important to first define what is meant by an interpretable AI/ML model. Many AI/ML methods are extraordinarily flexible and complex, meaning that they can approximate a wide range of functions and behaviors. These methods can also be difficult for humans to grasp the logic AI/ML methods used to arrive at their conclusions. As a result, many AI/ML methods are considered "black boxes" whose inner workings are largely incomprehensible to human users. Interpretable AI/ML leverages existing context and knowledge from human users to provide explanations that are (at the very least) more understandable than their black box counterparts (Samek, Wiegand, & Muller, 2017). As a result, the efficacy of a given interpretable AI method depends critically on the knowledge human users possess, which is used to contextualize AI logic.

Many traditional statistical models are inherently more interpretable than ML models because of constraints practitioners commonly impose on model complexity (Molnar, Casalicchio, & Bischl, 2020). In particular, traditional statistical models frequently make specific assumptions over functional forms and distributions (up to parameter values). Since these models are simpler and more constrained, it is generally easier for human analysts to correctly interpret model logic and results. Some models lend themselves particularly well to interpretation, including linear regression, decision trees, and decision rules (Huysmans, Dejaeger, Mues, Vanthienen, & Baesens, 2011). For these types of interpretable models, practitioners can often gain significant insights into the model by investigating a relatively small number of model components (e.g., parameter estimates, decision logic, goodness-of-fit measures). That said, more complex forms of these models (e.g., when the number of regressors is large) can still be difficult to interpret. Dimensionality reduction techniques (e.g., ridge regression, LASSO) can be useful in limiting the number of components practitioners must investigate to adequately interpret a given model (Tibshirani, 1996). Finally, even for more complex models, important interpretations can be drawn from a small number of model components. For example, in random forest models, just two components can be sufficient for analyzing tree structure and feature importance (Breiman L. , 2001).

Researchers have also developed and successfully used several model-agnostic methods of increasing ML interpretability. These methods are often categorized as either local, with the ability to explain specific model predictions, or global, with the ability to explain model behavior generally over a range of environments and conditions (Linardatos, Papastefanopoulos, & Kotsiantis, 2020). Counterfactual analyses are frequently used for local interpretability by permuting model inputs and assessing whether the associated change in predictions aligns with analysts' intuition and understanding of the system (Miller, 2019). Shapley values, a concept from cooperative game theory that fairly distributes payouts among players based on their contributions, can also provide local interpretability by quantifying how much each model input/feature contributes to forming a specific prediction (Strumbelj & Kononenko, 2014). Global interpretability methods are very similar,

quantifying either the importance of specific features or the effect of permuting features. Feature importance methods investigate shares of outcome variance explained by individual features or quantify the effect of removing features (where larger impacts to predictions indicate greater importance of features). Feature effects methods investigate the impact of augmenting a subset of model features and assessing whether effects on predictions matches analysts' understanding (Molnar, Casalicchio, & Bischl, 2020).

Finally, surrogate models are another model-agnostic method of increasing ML interpretability. Surrogate models have been used extensively in the field of uncertainty quantification, where ML models are used to approximate (i.e., act as a surrogate for) computer simulations or real-world experiments that are time- and resource-intensive to conduct. In that context, surrogate models can dramatically reduce time and cost of analysis, at the downside of increasing uncertainty in predictions (Sudret, Marelli, & Wiart, 2017). Surrogate models for interpretable ML are used to approximate black box algorithms, and the surrogate model form is chosen to be more interpretable than the black box algorithm it is approximating. As a result, the surrogate model behaves similarly to the original ML model but is more digestible and comprehensible to humans. This surrogate model approach often leverages traditional statistical models that tend to be inherently more interpretable, such as linear regression (including generalized linear models and generalized additive models), logistic regression, decision trees, and decision rules (Molnar, Interpretable machine learning, 2020).

4. REINFORCEMENT LEARNING

Reinforcement learning (RL) constitutes a set of ML algorithms that train an agent to learn to act in an environment in a way that maximizes some long-term reward function (Wang, et al., 2022). RL has been successfully utilized in a wide variety of applications from controlling autonomous vehicles to beating top human players in games such as Go (Holcomb, Porter, Ault, Mao, & Wang, 2018), and constitutes a key component of many artificial intelligence (AI) system. Ultimately, the objective of RL is to determine a *policy,* which can be loosely described as a set of rules that define how an agent should respond to different environment states. More precisely, a policy is a mapping of environment states to optimal actions. A key feature of RL is that a policy's actions maximize long-term rewards, not myopic rewards. Therefore, a successfully trained agent will sacrifice a smaller immediate reward in favor of a larger, long-term reward. While RL-based algorithms have been in existence for decades, the integration of deep neural networks (DNNs) into RL frameworks has greatly accelerated performance gains. Deep reinforcement learning (DRL) utilizes DNNs to identify features of complex environments and to encode environment knowledge in a way that otherwise would not be tractable to store computationally. RL algorithms can be broadly categorized according to two different attributes. The first is whether an algorithm is *on-policy* or *off-policy*. An on-policy algorithm interacts with the environment (i.e., collects data) using the most current learned policy; whereas, an off-policy algorithm relies on stored environment data that has been generated from past evolutions of the policy. The second category is whether the algorithm is *value-based* or *policy-based*. A value-based algorithm predicts the "goodness" (i.e., the expected total future discounted reward value) of each action in a given state and uses this goodness-value to select the best action. In contrast policy-based algorithms directly generates an action given the observed environment state. Many extensions of single-agent RL exist and consume much current research focus. Multi-agent RL considers algorithms that enable a team of agents to achieve a goal under various assumptions of information sharing and coordination. Adversarial RL seeks to train opposing agents against one another to achieve their respective goals.

The application of RL to a complex problem is not without challenges. A key challenge in any RL problem is the so-called *credit assignment problem*. Given that hundreds, or even thousands, of actions are taken during the course of an environment episode, how does one ascertain the degree to which a single action effected the final outcome? An extreme example of this would be a Chess game, in which the agent receives a terminal reward of "win" or "lose" but must allocate this reward over every past move. Another challenge is for applications that require a hierarchy of decisions spanning disparate time horizons. For example, one could consider an autonomous vehicle application that requires a high-level decision of planning the vehicles next way point seconds or minutes ahead but also requires lower-level immediate control inputs to keep the vehicle within its lane and to avoid collisions. In practice it is typically not possible to train a single RL agent to co-determine these higher- and lower-level decisions. Other challenges for RL include devising approaches to incorporate explainability of decisions as well as providing guarantees that the agent will not make unsafe decisions in safety-critical applications.

4.1. Hierarchical Reinforcement Learning

In this section, we provide a detailed overview of hierarchical reinforcement learning (HRL). HRL is a class of techniques developed for learning policies in complex applications where a hierarchy of decisions is required. Consider managing a fleet of vehicles to manage real time delivery demands. Here the higher-level problem is to assign a delivery to a vehicle based on the vehicle's current location and task queue as well as the locations and taskings of all other vehicles. The lower-level

problem is to prioritize the delivery taskings of individual vehicles and to determine the route by which they are delivered. HRL avoids the difficulty of training a single agent to make both these decisions by training separate agents to specialize in decisions at each level of the hierarchy. In general an HRL policy is comprised of two major components (Pateria, Subagdja, Tan, & Quek, 2021). The first is the hierarchical policy, which is a state-to-subtask-to-action mapping. In other words, the hierarchical policy is a trace across all levels of the hierarchy that ends in a primitive action taken by the agent. The second component is subtask discovery. This is the ability to identify how to partition each level of the hierarchy into separate task regions. For example, a person driving a vehicle may have different mental operating modes for driving on a smooth highway, merging onto an interstate, and handling stop-and-go traffic. The partitioning of the overall driving experience into these different operating modes is akin to discovering the subtasks which must be performed.

There are three broad categorizations for classifying HRL algorithms. The first is *subtask discovery;* that is, whether the algorithm automatically identifies and defines subtasks to be accomplished. If prior knowledge of the task structure exists, subtasks can be defined *a priori* by human experts and trained on their respective goals. The higher-level agent then must be trained to enable the appropriate subtask. On the other hand, algorithms can automatically define subtasks as they create the hierarchical policy. Regardless of whether subtasks are defined *a priori*, learning a hierarchical policy is nontrivial. The two other categorizations are whether an algorithm for single or multiple agents and whether the goal is to learn a single or multiple tasks.

Single-agent HRL algorithms have been demonstrated in a variety of complex applications. Tessler, et al. demonstrate the capacity of HRL to create a lifelong learning system that learns skills and retains knowledge that can be transferred between different tasks to play the computer game Minecraft (Tessler, Givony, Zahavy, Makowitz, & Mannor, 2017). They couple a Deep Skill Module, that has been trained *a priori* on various tasks, with a Hierarchical Deep Reinforcement Learning architecture that selects to either execute a single primitive action for a single time period or an entire skill over multiple time periods. The authors demonstrate their framework using a three-room Minecraft environment that requires three different tasks involving a block. Gu, et al. apply HRL to autonomous driving in a way the provides guarantees that the vehicle will not enter unsafe states (Gu, et al., 2023). Here they use an HRL framework with two levels. The high-level agent generates safe goals for the vehicle to navigate towards while the low-level agent navigates between adjacent goals. A noteworthy outcome of their research is a proof-of-concept that HRL-based schemes can ensure an agent does not execute a sequence of actions that ultimately leads to an undesirable outcome.

Applications for multi-agent HRL approaches have also been explored. Jendoubi and Bouffard use an options framework, where an upper-level agent selects amongst lower-level policies that are executed until their respective termination conditions are met (Jendoubi & Bouffard, 2023). They demonstrate their approach on two different scheduling problems for microgrids that require the coordination of multiple power- and load-generating components. The problem of air traffic management can also be addressed using multi-agent HRL (Spatharis, et al., 2023). Here agents are individual aircrafts which must coordinate their departure delays and trajectories in a way that satisfies airspace capacity constraints. A set of hierarchical policies operates at different levels of temporal abstraction to coordinate the aircraft. Lastly, HRL is integrated with graph neural networks by Yang for traffic signal control (Yang, 2023).

4.2. Reinforcement Learning Requirements and Outcomes

In this section, we describe the general computational and data requirements for implementing RL as well as the outcomes of the learning process. The computational requirements of RL vary significantly based on the complexity of the application. Classical RL algorithms can be implemented using simple data tables and elementary computers. Most modern RL applications involve Deep Reinforcement Learning (DRL), where RL is integrated with DNNs, which are computationally expensive. A DNN designed to discern complex environment features may be composed of hundreds to millions of neurons, which all must be trained using the backpropagation algorithm. Training DNNs efficiently requires Graphical Processing Units (GPUs), which can more efficiently perform the matrix operations required for backpropagation than a traditional Central Processing Unit (CPU). Larger neural networks are typically trained on GPU clusters, which contain tens to thousands of GPUs. In addition to neural networks, RL requires substantial computation to generate training data by conducting numerous environment simulations. In practice, executing these simulations, which are typically performed on CPUs, is often the bottleneck in the training process. Therefore, computational implementations often involve parallel processing with high-performance CPUs.

Unlike supervised machine learning techniques that use labeled training data, RL uses episodic data collected from interactions with an environment. These episodic data elements are composed of four components: (i) the current environment state, (ii) the action taken by the agent, (iii) the reward received, and (iv) the next environment state at which the agent arrives after taking the action. Off-policy RL algorithms can be trained using static, historical episodic data. For example, an agent could, in principle, be trained to play chess effectively if a large database of past games were available. However, whether using an off-policy or on-policy algorithm, an agent typically needs to periodically generate additional data by interacting with an environment as it continues to learn and improve its policy. This requires an accurate simulation of the environment so that the agent can observe realistic outcomes of its actions. Therefore, a critical component of RL is the verification and validation (V&V) of the underlying simulation environment used for generating training data. For a military application, the V&V required for this simulation is analogous to the vetting of wargaming or training scenarios for personnel.

The outcome of the RL process is a policy that prescribes the best action to take given the current state of the environment. Assuming properties of the environment remain static, an agent only needs to be trained once. In practice, most implementations of RL utilize DNNs, and the tangible outcome is a trained neural network that can output an optimal action given an encoded observation of the environment as an input. The energy requirements for evaluating a policy are the computations required to evaluate the DNN. In many real-world applications, the environment changes over time. In such cases, the agent must be periodically retrained using simulations that are more representative of the current environment. Fine tuning an agent for differing environment conditions is often much faster than the time required to train an agent from scratch. Much of the knowledge acquired by the agent can be "transferred" to similar tasks. For example, an RL agent trained to drive in good weather could utilize much of its knowledge of object recognition and vehicle dynamics to learn to drive in inclement weather.

4.3. Explainability Approaches for Reinforcement Learning

Explainable Artificial Intelligence (XAI) is a rapidly evolving area of research. The goal is to develop techniques that provide human-understandable explanations of what input features are most salient to an AI-systems' decisions and the impacts those features have. Currently there are no universally

agreed-on metrics for defining explainability, and the XAI approaches taken tend to be very specific to their intended audiences. The consequence is that XAI approaches are relatively inchoate and not standardized, so it is not straightforward to pinpoint the best techniques to utilize for a given application.

A comprehensive overview of XAI approaches for DRL is given by Heuillet, et al. (Heuillet, Couthouis, & Diaz-Rodriguez, 2021). The authors detail numerous approaches for XAI that have been considered to date for DRL. State representation learning is class of approaches for building a low-dimensional representation of a complex state space so that meaningful features can be identified. Several techniques attempt to learn explainability while learning the agent's policy. Reward decomposition divides an agent's reward function into different parts so that actions can be classified by the reward components they are intended to maximize. It is also possible to obtain minimal sufficient explanations, the smallest set of reasons why an agent takes a particular action, as well as construct action influence models that trace relations between actions and outcomes. An approach specific to HRL is observing which subgoals the high-level agent determines are optimal at a given time based on environment features. Some approaches utilize those developed for DNN image classification, such as projecting saliency maps (i.e., heat maps) onto a visual input to highlight the features that most influenced the DNN's output.

That explainability only has meaning with respect to human beliefs and interpretations is explored more deeply by Vouros (Vouros, 2022). Here different models for the explainability problem are proposed that include various criteria for how humans may interpretate an agent's actions with respect to their understanding of the environment as well as the agent's objectives and abilities. Some more general approaches for XAI are outlined such as identifying critical state-action pairs, where taking the wrong action causes a large decrease in future rewards. A related approach involves constructing contrastive explanations by choosing a different action than that prescribed by the optimal policy. Off-policy evaluation identifies influential environment state transitions by estimating the value of a policy using data collected from a different policy.

5. ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING APPLICATIONS IN PRIVATE INDUSTRY AND GOVERNMENT

AI and ML have become central to technological advancements across both private industry and government sectors. The ability to analyze vast amounts of data, automate processes, and improve decision-making efficiency has made AI/ML invaluable tools in a wide range of applications. While the private sector has leveraged AI/ML to enhance operations, logistics, manufacturing, and customer interactions, the public sector has focused on using these technologies to improve public services, increase safety, and streamline complex systems such as defense and transportation. Despite the broad adoption of AI/ML, challenges such as data integrity, human-computer interaction, and system transparency remain, particularly in high-consequence environments. This section explores the diverse applications of AI/ML in both sectors, highlighting their impact and the unique obstacles faced in these different contexts.

5.1. Overview of AI/ML Applications in Government and Industry

Artificial Intelligence (AI) and Machine Learning (ML) have emerged as pivotal technologies across both government and private industry, offering substantial potential to revolutionize operations and decision-making processes. In government, these technologies are being applied to critical areas such as defense, public safety, healthcare, and infrastructure management, while private industry uses AI/ML to optimize processes, enhance customer experiences, and increase profitability. The integration of AI/ML has led to considerable successes in both sectors, with improved efficiency, precision, and the ability to analyze vast amounts of data more effectively than traditional methods.

Despite these successes, the adoption of AI/ML also presents significant challenges, particularly in areas requiring high-consequence decision-making, where errors or system failures could result in severe harm. Issues such as data integrity, algorithmic bias, ethical concerns, and transparency often complicate the deployment of AI/ML systems, especially when human oversight is limited or when these technologies are applied in sensitive environments like healthcare or defense. This section explores several real-world applications of AI/ML across both sectors, examining their strengths, the opportunities they present, and the obstacles they face.

One key area of AI/ML application is in government-run traffic management systems, where cities like Los Angeles have leveraged AI to reduce congestion and enhance public safety. For example, Los Angeles employs an advanced adaptive traffic control system that uses real-time data from thousands of cameras and sensors to adjust traffic signals. This AI-powered system has been shown to reduce congestion by up to 16% during peak hours. Similar systems have been implemented in other cities around the world, such as Hangzhou, China, where Alibaba's City Brain uses AI to analyze data from millions of sensors, optimizing traffic flow and reducing emergency response times by up to 50% (PYMNTS, 2022). These systems are effective in real-time adjustments to traffic flows, providing a clear efficiency boost compared to human-operated systems. However, challenges such as data quality and potential biases in how traffic is prioritized can lead to uneven outcomes. For instance, flawed sensor data or biased algorithmic designs could disproportionately affect certain communities, highlighting the importance of accurate, unbiased data in AI-driven public systems.

In the defense sector, one of the most significant AI initiatives is the U.S. Department of Defense's Project Maven. This project utilizes machine learning algorithms to analyze video data collected by drones, helping military personnel quickly identify and track objects of interest. The system greatly enhances the speed and accuracy of intelligence gathering, allowing for faster decision-making in combat scenarios. The key strength here is the efficiency AI brings to intelligence operations by

sifting through massive amounts of data far more quickly than human analysts. However, this efficiency comes with ethical concerns. Critics have raised questions about the potential for AI-driven military operations to make decisions that could result in unintended civilian casualties. The "black box" nature of some machine learning algorithms further complicates matters, as the lack of transparency makes it difficult for humans to understand the rationale behind AI-driven decisions. This raises concerns about accountability and the potential for AI to be used in lethal operations without proper oversight (Cummings, 2023).

Another critical defense application is the use of AI in autonomous systems, such as unmanned aerial vehicles (UAVs) and autonomous combat drones. These systems use AI to navigate, identify targets, and execute missions with minimal human intervention. For example, the U.S. military has been testing AI-powered autonomous drones capable of conducting reconnaissance and even engaging in combat scenarios. The strength of these autonomous systems lies in their ability to operate in hazardous environments without risking human lives, improving operational efficiency and reducing personnel exposure to danger. However, they raise significant challenges related to accountability, control, and ethical decision-making. Critics argue that delegating life-or-death decisions to machines, particularly in unpredictable combat environments, is fraught with risks, and the lack of human oversight could lead to unintended consequences, including violations of international law (RAND, 2020).

Healthcare offers another promising but complex domain for AI/ML, particularly in improving diagnostics and personalized medicine. IBM's Watson for Oncology, for instance, uses AI to analyze patient data and medical literature to assist doctors in determining cancer treatments. While this application of AI can lead to more accurate and personalized care, challenges such as trust in machine-generated recommendations and the risk of bias in the data used to train these models remain significant obstacles. If these biases are not addressed, AI-driven healthcare systems may unintentionally perpetuate inequalities in care (Ross & Swetlitz, 2017).

In private industry, AI/ML has been particularly successful in financial services, where companies use these technologies to detect and prevent financial crimes such as fraud and money laundering. For example, machine learning is being increasingly used in anti-money laundering (AML) efforts to analyze large datasets, recognize suspicious patterns, and flag high-risk transactions for further investigation. According to a McKinsey report, machine learning has become a game-changer in the fight against money laundering by improving the efficiency and accuracy of detection systems. These AI-powered systems significantly reduce false positives and help institutions comply with regulatory requirements more effectively. However, challenges remain, particularly in the need to continuously update these models to keep pace with evolving criminal tactics. Additionally, maintaining the transparency and interpretability of these AI systems is critical to ensure that financial institutions and regulators can trust the decisions being made by the models (Doppalapudi, et al., 2022).

In summary, AI and ML have demonstrated the capacity to transform both government and private industry by increasing efficiency, enhancing decision-making, and providing innovative solutions to complex problems. Government applications such as traffic management and defense illustrate how AI can streamline operations and improve safety, while healthcare applications highlight the potential for improved diagnostics and personalized care. In private industry, the use of AI/ML in areas like financial fraud detection shows the value of these technologies in mitigating risks and optimizing operations. However, these advancements are tempered by challenges such as data quality, algorithmic bias, and ethical considerations, which must be addressed to ensure the responsible and effective use of AI/ML. As these technologies continue to evolve, finding the right balance between innovation and oversight will be critical to their successful deployment. The next

section will delve deeper into how AI/ML is applied in high-consequence systems, where the stakes are exceptionally high and reliability is paramount.

5.2. Planes, Train(ing)s, and Automobiles: Applications in High-Consequence Systems

High-consequence systems are those in which failures or errors can lead to significant harm or substantial negative outcomes, affecting safety, security, or critical operations. These systems often operate in environments where reliability and accuracy are paramount, as the stakes involved are exceptionally high. One notable example is self-driving vehicles, which not only have to process large amounts of information and make real-time choices on the road but also integrate into systems with human-driven vehicles. This application is generally seen as a "high-consequence" application of AI/ML because errors made on the road can cause harm to drivers, passengers, pedestrians, and property (Goodall, 2014). Additionally, these systems must be able to process and react to a large number of possible scenarios, each of which may be novel to the system. Some of these problems can be solved by generalizing from prior experience, while others may require novel approaches. For example, a system trained to avoid collisions with adult humans by stopping will likely be able to handle a novel scenario like a child in the road with the same solution—stopping the car. However, if a car detects darker pavement in the winter and assumes it's water when it's actually black ice, the car's solution—braking or slowing down—may not prevent the negative outcome of skidding or losing control.

Currently, a solution to this challenge is to have a person in the vehicle who is ready to take control if the autonomous driving system encounters something novel or reacts incorrectly (Lin, et al., 2021). Human oversight is framed as a way to reduce the likelihood of negative outcomes and potentially serve as training data to teach the system how to respond better in the future. This hybrid system of human-in-the-loop (HITL) or human-on-the-loop (HOTL) is common in AI/ML applications where the consequences of errors are severe (Cummings, 2023). HITL systems involve active human participation in the decision-making process, while HOTL systems place humans in a supervisory role, monitoring the AI's actions and intervening when necessary.

Questions have been raised about the feasibility of applying these systems to other high-consequence environments like air traffic control and space applications. Air traffic control, which relies on a variety of information interfaces and the complex orchestration of assets, is often cited as a field that cannot easily be replaced by AI/ML systems. While AI can assist in managing routine tasks, the unpredictability of human behavior and the potential for system failures make it critical that humans remain involved in these systems.

Moreover, challenges such as explainability and interpretability further complicate the deployment of AI/ML in these high-stakes contexts (Doshi-Velez & Kim, 2017). Humans need to understand the rationale behind AI decisions to trust the system and intervene appropriately. In environments like air traffic control, this need for explainability means that AI/ML systems must provide not only accurate decisions but also transparent reasoning that can be quickly understood by human operators. Although AI systems can operate faster and more efficiently than humans in some scenarios, the need for oversight in high-consequence environments ensures that human involvement will continue to be necessary for the foreseeable future (Ribeiro, Singh, & Guestrin, 2016).

6. CONCLUSION

AI and ML have transformed both private industry and government, significantly enhancing efficiency, automation, and decision-making. These technologies are now fundamental in sectors like manufacturing, logistics, healthcare, and defense, enabling organizations to handle vast amounts of data and improve operational outcomes. In private industry, AI/ML is optimizing supply chains, predicting failures, and supporting customer engagement. Meanwhile, government applications focus on improving public services, transportation, defense, and public safety.

Despite these benefits, challenges persist. Ensuring data integrity and handling missing or incomplete data are crucial for maintaining model accuracy, especially in high-consequence environments like autonomous vehicles and air traffic control, where the stakes are high. Moreover, AI/ML models often lack transparency, making it difficult for humans to interpret decisions, which is vital in critical settings. The rise of explainable AI (XAI) is beginning to address this issue, but further advancements are needed to build trust and ensure accountability.

The human-computer interface remains essential in AI/ML systems, particularly in high-stakes applications where human oversight helps prevent catastrophic outcomes. HITL and HOTL systems allow humans to guide or supervise AI actions, balancing the efficiency of AI with the need for human intervention. As AI becomes more autonomous, maintaining this balance—without over-relying on AI or slowing down operations—will be crucial for safety and trust.

Looking ahead, further research, development, and policy will be needed to overcome these challenges. Ensuring AI/ML systems are transparent, reliable, and ethically deployed is vital for their continued success. AI and ML hold vast potential to revolutionize industries and improve decision-making, but their future hinges on responsible development that aligns automation with human oversight, safety, and ethical considerations.

7. REFERENCES

Bansal, G., Tongshuang, W., Zhu, J., Fok, R., Nushi, B., Kamar, E., . . . Weld, D. S. (2021). Does the Whole Exceed its Parts? The Effect of AI Explanations in Complementary Team Performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

Baraldi, A., & Enders, C. (2010). An introduction to modern missing data analyses. *Journal of school psychology*, 5-37.

Boulesteix, A.-L., & Schmid, M. (2014). Machine learning versus statistical modeling. *Biometrical Journal*, 588-593.

Breiman, L. (2001). Random forests. *Machine Learning*.

Breiman, L. (2001). Statistical modeling: Two cultures. *Statistical Science*, 199-231.

Brose, C. (2020). *The Kill Chain: Defending America in the Future of High-Tech Warfare*. Hachette Books.

Bucinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 1-21.

Chong, L., Zhang, G., Goucher-Lambert, K., Kotovsky, K., & Cagan, J. (2022). Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on the adoption of AI advice. *Computers in Human Behavior*.

Collins, L., Schafer, J., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological methods*, 330.

Cummings, M. (2023). Revising human-systems engineering principles for embedded AI applications. *Frontiers in Neuroergonomics*.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society: Series B (Methodological)*, 1-22.

Doppalapudi, P., Kumar, P., Murphy, A., Werner, S., Zhang, S., Rougeaux, C., & Stearns, R. (2022). The fight against money laundering: Machine learning is a game changer. *McKinsey & Company*.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint*.

EU Commission. (2020). On Artificial Intelligence--A European Approach to Excellence and Trust. *EU Commission White Paper*.

Fahle, S., Prinz, C. P., & Kuhlenkotter, B. (2020). Systematic review on machine learning (ML) methods for manufacturing processes - Identifying artificial intelligence (AI) methods for field application. *Procedia CIRP 93*, 413-418.

Fitzsimons, G. J., & Lehmann, D. R. (2004). Reactance to recommendations: When unsolicited advice yields contrary responses. *Marketing Science*, 82-94.

Goodall, N. (2014). Ethical decision making during automated vehicle crashes. *Transportation Research Record*, 58-65.

Green, B., & Chen, Y. (2019). The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 1-24.

Gu, Z., Gao, L., Ma, H., Li, S., Zheng, S., Jing, W., & Chen, J. (2023). Safe-state enhancement method for autonomous driving via direct hierarchical reinforcement learning. *IEEE Transactions on Intelligent Transportation Systems*, 9966-9983.

Gupta, S., Modgil, S., Bhattacharyya, S., & Bose, I. (2022). Artificial intelligence for decision support systems in the field of operations research: Review and future scope of research. *Artificial Intelligence in Operations Management*, 215-274.

Heuillet, A., Couthouis, F., & Diaz-Rodriguez, N. (2021). Explainability in deep reinforcement learning. *Knowledge-Based Systems*.

Holcomb, S., Porter, W., Ault, S., Mao, G., & Wang, J. (2018). Overview on deepmind and its AlphaGo Zero AI. *Proceedings of the 2018 international conference on big data and education*, 67-71.

Holsapple, C. (2008). DSS architecture and types. In F. Burnstein, & C. Holsapple, *Handbook on decision support systems 1: Basic themes* (pp. 163-189). New York: Springer.

Huysmans, J., Dejaeger, K., Mues, C., Vanthienen, J., & Baesens, B. (2011). An empirical evaluation of the comprehensibility of decision table, tree, and rule based predictive models. *Decision Support Systems*, 141-154.

Jacovi, A., Marasovic, A., Miller, T., & Goldberg, Y. (2021). Formalizing Trust in Artificial Intelligence: Prerequisites, Causes and Goals of Human Trust in AI. *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 624-635.

Jendoubi, I., & Bouffard, F. (2023). Multi-agent hierarchical reinforcement learning for energy management. *Applied Energy*.

Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). Trends in big data analytics. *Journal of parallel and distributed computing*, 2561-2573.

Lin, B., Lin, K., Lin, C., Lu, Y., Huang, Z., & Chen, X. (2021). Computation offloading strategy based on deep reinforcement learning for connected and autonomous vehicle in vehicular edge computing. *Journal of Cloud Computing*, 33.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A review of machine learning interpretability methods. *Entropy*.

Litte, R., & Rubin, D. (2002). *Statistical analysis with missing data.* John Wiley & Sons.

Little, R. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 125-134.

Miller, T. (2019). Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, 1-38.

Molenberghs, G., Beunckens, C., Sotto, C., & Kenward, M. (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *Journal of the Royal Statistical Society: Series B (Methodology)*, 371-388.

Molnar, C. (2020). *Interpretable machine learning.* Lulu.com.

Molnar, C., Casalicchio, G., & Bischl, B. (2020). Interpretable Machine Learning - A Brief History, State-of-the-Art, and Challenges. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 417-431.

Morse, P., & Kimball, G. (1947). Operational Research in the British Army 1939-1945. *UK National Archives*.

OECD. (2016). *Artificial intelligence on society.* Paris: OECD Publishing. Retrieved from OECD Publishing.

Park, J. S., Barber, R., Kirlik, A., & Karahalios, K. (2019). A Slow Algorithm Improves Users' Assessments of the Algorithm's Accuracy. *Proceedings of the ACM on Human-Computer Interaction*, 1-15.

Pateria, S., Subagdja, B., Tan, A., & Quek, C. (2021). Hierarchical reinforcement learning: A comprehensie survey. *ACM Computing Surveys*, 1-35.

Pournader, M., Ghaderi, H., Hassanzadegan, A., & Fahimnia, B. (2021). Artificial intelligence applications in supply chain management. *International Journal of Production Economics*.

PYMNTS. (2022, February). *PYMNTS*. Retrieved from Cities Turn to AI-Powered Solutions to Tackle Traffic Challenges: https://www.pymnts.com/connectedeconomy/2022/cities-turn-to-ai-powered-solutions-to-tackle-traffic-challenges/

RAND. (2020). *Military applications of artificial intellgience.* Santa Monica: RAND Corporation.

Ribeiro, M., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of a classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*.

Ross, C., & Swetlitz, I. (2017, September). *Stat News*. Retrieved from IBM pitched its Watson supercomputer as a revolution in cancer care. It's nowhere close.: https://www.statnews.com/2017/09/05/watson-ibm-cancer/

Rubin, D. (1988). An overview of multiple imputation. *Proceedings of the survey research methods section of the American statistical association*, 84.

Samek, W., Wiegand, T., & Muller, K.-R. (2017). Explainable Artificial Intelligence: Understanding, Visualizing, and Interpreting Deep Learning Models. *arXiv Preprint*.

Schafer, J., & Graham, J. (2002). Missing data: Our view of the state of the art. *Psychological methods*, 147.

Schaffer, J., O'Donovan, J., Michaelis, J., Raglin, A., & Hollerer, T. (2019). I can do better than your AI: Expertise and explanations. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 240-251.

Spatharis, C., Bastas, A., Kravaris, T., Blekas, K., Vouros, G., & Cordero, J. (2023). Hierarchical multiagent reinforcement learning schemes for air traffic management. *Neural Computing and Applications*, 1-13.

Sprague Jr., R. (1980). A framework for the development of decision support systems. *MIS quaterly*, 1-26.

Strumbelj, E., & Kononenko, I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 647-665.

Sudret, B., Marelli, S., & Wiart, J. (2017). Surrogate models for uncertainty quantification: An overview. *2017 11th European conference on antennas and propagation*, 793-797.

Tessler, C., Givony, S., Zahavy, T., Makowitz, D., & Mannor, S. (2017). A deep hierarchical approach to lifelong learning in Minecraft. *Proceedings of the AAAI conference on artificial intelligence*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 267-288.

Vouros, G. (2022). Explainable deep reinforcement learning: State of the art and challenges. *ACM Computing Surveys*, 1-39.

Wacker, J. (1998). A definition of theory: Research guidelines for different theory-building research methods in operations management. *Journal of Operations Management*, 361-385.

Wang, X., Wang, S., Liang, X., Zhao, D., Huang, J., Xu, X., . . . Miao, Q. (2022). Deep reinforcement learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 5064-5078.

Woschank, M., Rauch, E., & Zsifkovits, H. (2020). A Review of Futher Directions for Artificial Intelligence, Machine Learning, and Deep Learning in Smart Logistics. *Sustainability*.

Yang, S. (2023). Hierarchical graph multi-agent reinforcement learning for traffic signal control. *Information Sciences*, 55-72.

Zerilli, J., Bhatt, U., & Weller, A. (2022). How transparency modulates trust in artificial intelligence. *Patterns*.

DISTRIBUTION

Email—Internal

Name	Org.	Sandia Email Address
Technical Library	1911	sanddocs@sandia.gov

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.