Neural Networks and Nervous Wrecks

Title: 20250508_NN_NW_Promise_Pitfalls/pdf
Date: 8th May 2025 Version: 1.0

Neural Networks and Nervous Wrecks: What AI Really Is—Its History, Promise, and Pitfalls

Novalytics Gibraltar

Abstract—Artificial Intelligence (AI) has rapidly evolved from theoretical constructs to a powerful, practical force driving innovation across diverse sectors. This paper provides a comprehensive overview of AI's historical development, technical foundations, applications, and limitations. It explores the transition from symbolic reasoning and expert systems to modern machine learning and deep learning paradigms. Key concepts such as neural networks, supervised and reinforcement learning, model training, and AI infrastructure are explained, with visual diagrams to aid understanding. The paper also discusses the real-world impact of AI in IT, finance, and scientific research, highlighting its strengths in pattern recognition, big data analysis, automation, and forecasting. At the same time, critical limitations are examined, including lack of contextual understanding, data bias, interpretability issues, and overreliance risks. The conclusion offers guidance on when and how to responsibly apply AI, emphasising hybrid approaches and the need for human oversight in high-stakes settings. Overall, the work aims to equip readers with a grounded, nuanced understanding of AI as a transformative yet bounded technology.

Keywords—*Artificial Intelligence, Machine Learning, Neural Networks, Supervised Learning, Reinforcement Learning, AI Limitations, AI Ethics, Deep Learning, Automation, Pattern Recognition, Information Technology, Finance, Scientific Research*

Contents

1. Introduction

Artificial intelligence (AI) refers to machines or software that perform tasks traditionally requiring human intelligence, such as learning, problem solving, decision making, and language comprehension. Today, AI systems are integrated into various aspects of life, from virtual assistants and recommendation algorithms to autonomous vehicles and advanced analytics. With the advancement of AI capabilities, discussions surrounding its potential and limitations have intensified, including questions about whether AI will exceed human intelligence, its transformative impact on industries, and associated risks [1]. To fully understand these concerns, it is crucial to explore *the history of AI*, *the mechanisms behind its functionality*, and *its successes and shortcomings*.

This document presents a detailed overview of AI's evolution and its current applications. It begins by examining the history of AI, highlighting significant milestones, from early symbolic reasoning programmes [2] to modern deep learning and generative models [3]. The following section explains the technical foundations of AI, exploring core concepts such as algorithms, neural networks, and various learning paradigms (supervised, unsupervised, and reinforcement learning). Subsequently, the discussion turns to the influence of AI on contemporary society, especially in the fields of information technology, finance, and data-driven research, showcasing its transformative role [6]. The paper also identifies key domains where AI excels, using examples from IT operations, financial analysis, and data science applications [4]. Furthermore, we address areas where AI struggle, including issues related to generalisation, contextual understanding, data dependence, bias, and the opaque nature of many AI models [5]. The conclusion highlights situations where AI may not be the optimal solution, emphasising the risks of overreliance and advocating traditional methods or human judgment when appropriate [9].

The objective is to provide a comprehensive, up-to-date understanding of AI's development, functioning, applications, and limitations, thereby laying the groundwork for thoughtful, responsible integration of AI across various sectors.

2. History of AI

The concept of intelligent machines has been explored for centuries in both science and fiction, but artificial intelligence (AI) as a formal research field emerged in the mid-20th century. Pioneering work by Alan Turing in the 1930s and 1940s on computability and the concept of a "Turing machine" laid the theoretical foundations for understanding machine intelligence [1]. In 1950, Turing introduced the famous Turing test, proposing that a machine could be considered 'intelligent' if its conversational responses were indistinguishable from those of a human. Several years later, in 1956, the term "Artificial Intelligence" was coined at the Dartmouth Conference, organised by John McCarthy and colleagues, a pivotal event that is often regarded as the official launch of AI as a research field [2]. Early optimism was high; the researchers at Dartmouth, along with contemporaries such as Newell and Simon, believed that human-level AI could be achieved in a few decades.

2.1. The Era of Symbolic AI (1950s–1960s)

The 1950s and 1960s witnessed the rise of symbolic AI, which focused on explicit logical rules and symbolic representations of knowledge [3]. The prevailing idea was that human reasoning could be emulated by machines that manipulate symbols based on predefined rules. Notable early AI programmes include the logic Theorist (1956) and the General Problem Solver (GPS) (1957), both developed by Allen Newell and Herbert Simon. These programmes successfully proved mathematical theorems and solved puzzles using logical rules [6]. In 1958, Frank Rosenblatt introduced the perceptron, an early single-layer neural network capable of learning to classify simple patterns [8]. However, at this stage, perceptrons were still considered a part of the symbolic AI paradigm, which functions as linear classifiers with limited capabilities.

Early AI research also explored games and language processing as testing grounds for artificial intelligence. In 1952, Arthur Samuel developed a checkers-playing programme that could improve its performance through self-play – an early demonstration of machine learning principles [4]. In 1966, Joseph Weizenbaum's ELIZA programme showcased natural language interaction by engaging users in typed conversation using simple pattern matching scripts [9]. These projects captured the public imagination and demonstrated "intelligent" behaviour, but were largely limited by manual encoded rules or very simple learning mechanisms, highlighting the need for more robust learning and adaptability in AI systems [5].

2.2. Challenges and the First "AI Winter" (1970s)

By the late 1960s, the initial optimism surrounding AI began to give way to reality. Despite early successes, symbolic AI systems revealed significant limitations. They struggled to handle the ambiguity and complexity of real-world scenarios beyond their programmed rules [6]. As one contemporary critique put it, these systems were brittle, effective only within narrow, predefined contexts. Prominent researchers, such as Marvin Minsky, warned in 1969 that Rosenblatt's perceptron could not solve simple non-linear problems, such as the XOR problem, which dampened enthusiasm for neural approaches [1]. The grand promises of achieving human-level AI were, in hindsight, premature.

Consequently, the 1970s ushered in the first AI Winter – a period of reduced funding and waning interest in AI research [3]. Several factors contributed to this downturn [2]:

Computational limitations: The early computers lacked the processing power and memory required to support the ambitious AI programmes envisioned [5]. Complex reasoning and large knowledge bases quickly exhausted the available resources.

Knowledge acquisition bottleneck: Symbolic AI systems required extensive manual encoding of domain knowledge. Scaling this approach to capture the knowledge of the common sense proved infeasible [8].

Inability to learn and generalise: Rule-based systems could not learn from new data or adapt to unforeseen situations. They performed poorly outside the exact scenarios for which they were programmed [9].

Expectation backlash: The hype about AI in the 1960s led to unrealistic expectations. When progress stalled, investors and government agencies became increasingly sceptical [4].

As funding dried up, many AI projects were cancelled throughout the 1970s [6]. This era served as a harsh lesson for the AI community about the dangers of overpromising and underdelivering. However, important research continued in isolated pockets. For example, in 1975, computer scientist Ed Feigenbaum developed DENDRAL and later MYCIN, early expert systems for chemistry and medicine that would later serve as the foundation for a resurgence of AI through expert knowledge capture [7].

2.3. Expert Systems and the Second AI Winter (1980s)

AI research saw a resurgence in the early 1980s with the advent of expert systems. These programmes were designed to simulate the decision-making of human domain experts by encoding large sets of if-then rules. One notable example was XCON, an expert system developed by DEC in the late 1970s for configuring computer systems. Expert systems achieved some commercial success in fields such as medical diagnosis (e.g. MYCIN), geology (prospecting) and finance, demonstrating real economic value by capturing specialised knowledge [9]. This period is often referred to as the AI spring of the 1980s, with governments (such as Japan's Fifth Generation Computer project) and companies once again investing heavily in AI.

However, the limitations of expert systems soon became apparent. These systems were costly to build and maintain, as the "knowledge engineering" process of manually extracting rules from experts was both time-consuming and expensive. Furthermore, expert systems lacked the ability to learn. When the environment changed or exceptions occurred, their rigid structure caused failures. By the late 1980s, enthusiasm for AI began to wane once again. The market became saturated with weak expert systems, many of which failed to meet the high expectations set for them, leading to the second AI winter around 1987-1993 [3]. Many AI companies folded and research funding was once again reduced.

Despite these setbacks, the 1980s laid important groundwork for future AI developments. During this time, researchers began to explore connectionist approaches, notably neural networks, more deeply. In 1986, Rumelhart, Hinton, and Williams rediscovered and popularised the backpropagation algorithm for training multilayer neural networks, which overcame the earlier limitations of single-layer perceptrons [6]. Meanwhile, Judea Pearl developed probabilistic graphical models, also known as Bayesian networks, to reason under uncertainty [8]. Although these ideas did not fully yield results until later, they represented a significant shift from pure rule-based AI to methods capable of handling uncertainty and learning from data.

2.4. The Machine Learning Era (1990s–2000s)

By the 1990s, AI's focus shifted to machine learning, algorithms that enable computers to identify patterns and make predictions from data. Rather than manually encoding behaviour through rules, researchers began developing statistical techniques that could be trained on large datasets. This era saw the maturation of methods such as decision trees, support vector machines (SVMs), ensemble methods, and neural networks with a few hidden layers, sometimes referred to as "shallow" neural nets.

A significant milestone occurred in 1997 when IBM's Deep Blue chess computer defeated world champion Garry Kasparov, the first time a reigning world chess champion lost to a computer in a tournament setting [1]. Deep Blue's victory was achieved through brute-force search and expert-tuned heuristics, rather than through learning. However, it demonstrated the immense computational power available for solving complex problems. In the same year, Sepp Hochreiter and Jürgen Schmidhuber introduced the Long-Short-Term Memory (LSTM) neural network, which enabled machines to learn from sequential data and would later prove crucial for speech and language applications [3].

The late 1990s and early 2000s saw significant advancements, particularly in speech recognition and computer vision, through machine learning techniques. In 1989, Yann LeCun and colleagues demonstrated a convolutional neural network (CNN) capable of recognising handwritten characters – a precursor to modern image recognition [6]. By the 2000s, statistical approaches had fully taken over many AI fields, often rebranded as machine learning or data mining. A landmark achievement occurred in 2009, when Fei-Fei Li's team introduced ImageNet, a massive labelled image dataset that would become a catalyst for major progress in computer vision [8].

Another notable event was IBM's Watson system defeating human champions on the quiz show Jeopardy! in 2011 [9]. Watson combined machine learning, natural language processing, and information retrieval techniques to answer general knowledge questions, demonstrating the potential of AI to handle unstructured language and vast amounts of information under time pressure.

2.5. The Deep Learning Revolution (2010s)

The 2010s witnessed a dramatic resurgence of neural networks, now with many layers, marking the rise of deep learning. Three key factors converged to enable this revolution: (1) the availability of massive datasets, such as ImageNet's millions of images, (2) the emergence of more powerful hardware, particularly graphics processing units (GPUs) that accelerated neural network computations, and (3) the development of improved algorithms and architectures, thanks to researchers who persisted with neural approaches during previous lean years.

A defining breakthrough occurred in 2012, when a deep convolutional neural network (CNN) known as AlexNet, developed by Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, won the ImageNet image recognition competition by a significant margin, achieving far better accuracy than prior approaches [6]. AlexNet, with 8 learnt layers and trained on GPUs, demonstrated the power of deep learning as a general approach. Its success prompted an explosion of deep learning research and applications [1]. In the following years, deep neural networks dominated benchmarks in computer vision, speech recognition, and eventually natural language processing (NLP). For example, in 2014, Ian Goodfellow's invention of generative adversarial networks (GAN) introduced a novel way for neural networks to generate surprisingly realistic images [3].

Deep learning also enabled significant advances in games. In 2016, DeepMind's AlphaGo system, which combined deep neural networks with reinforcement learning, defeated the top Go player Lee Sedol, a feat previously considered at least a decade away [9]. Unlike chess, Go has a vastly larger search space, and AlphaGo's victory demonstrated the power of deep learning to handle complexity by learning value and policy networks from data, including both human expert games and self-play. AlphaGo's successors, such as AlphaZero and AlphaStar, have since mastered additional games and even solved complex problems like protein folding. In 2020, DeepMind AlphaFold achieved a breakthrough in the prediction of 3D protein structures, outperforming decades of prior research in that field [8].

A further paradigm shift occurred in 2017 with the introduction of the transformer architecture by Vaswani et al. in their paper Attention is All You Need [5]. Transformers enabled much larger and more effective neural networks for sequence data by replacing recurrent architectures with attention mechanisms that capture long-range dependencies. This development led to the era of large-scale language models. In 2018, OpenAI's GPT (Generative Pre-trained Transformer) demonstrated that a transformer-based network trained on massive text corpora could generate coherent text [9]. By 2020, OpenAI introduced GPT-3, a language model with a staggering 175 billion parameters, capable of performing a wide range of language tasks with minimal prompting [7]. The public release of ChatGPT at the end of 2022, a conversational interface built on GPT-3.5, captured the attention of the global public, as millions experienced an AI system capable of producing remarkably human-like dialogue on virtually any topic [1]. In 2023, OpenAI's GPT-4 and other competitors, such as Google's Bard, pushed the boundaries further with multimodal abilities (processing both images and text) and improved reasoning, though concerns about factual accuracy and misuse remain [3].

By the early 2020s, AI has indisputably moved from the laboratory to widespread deployment. AI techniques are integral to many everyday technologies (search engines, smartphones, vehicles), and AI research continues at an accelerated pace. We now turn to an explanation of how these AI systems actually work from a technical standpoint.

Year	Milestone	Significance
1950	Turing Test proposed	Defined a benchmark for evaluating machine intelligence.
1956	Dartmouth Conference	Term "Artificial Intelligence" coined; launch of AI as a formal discipline.
1958	Rosenblatt's Perceptron	Introduced the first neural network model capable of learning from data.
1966	ELIZA chatbot	Early demonstration of natural language processing simulating conversation.
1970s	First AI Winter	Research and funding declined due to limitations of symbolic AI approaches.
1980s	Expert systems boom	Deployment of rule-based systems in commercial and industrial settings.
1987	Second AI Winter	Decline in interest following failures of expert systems to scale or adapt.
1997	Deep Blue defeats Kasparov	First time a reigning world chess champion lost to a computer.
2006	ImageNet project begins	Enabled large-scale supervised learning in computer vision.
2012	AlexNet wins ImageNet	Demonstrated the power of deep learning for image recognition tasks.
2016	AlphaGo beats Lee Sedol	Breakthrough in deep reinforcement learning for complex strategic games.
2022	ChatGPT released	Transformer-based language models gain mainstream adoption.

Table 1. Abbreviated timeline of AI history highlighting select milestones. Citations for these milestones are available in the main text.

3. Technical Workings of AI Systems

Modern AI systems are built upon a combination of algorithms and computational architectures that allow data learning and the performance of complex tasks. In this section, we explain the core concepts of how AI functions, beginning with algorithms and progressing to the specialised models and learning paradigms that define today's AI.

3.1. Algorithms and Models in AI

At its core, an algorithm is a step-by-step procedure to solve a problem or perform a computation. In AI, algorithms often take the form of training procedures that adjust the parameters of a model to improve performance on a given task. A model in AI is the mathematical structure or programme that makes predictions or decisions. For example, a linear regression model is defined by a linear equation with specific coefficients. A training algorithm, such as the least-squares fit, finds values for these coefficients that best fit the training data. In more complex AI systems, the model might be a multilayer neural network with millions of parameters, and the training algorithm might be stochastic gradient descent.

Two broad families of AI models can be distinguished:

Symbolic models: These explicitly encode knowledge and logic. For example, the model of an expert system is a rule knowledge base, and an algorithm (such as a logical inference engine) operates on this knowledge base to derive conclusions.

Subsymbolic models: These include neural networks and other distributed representations, where knowledge is stored in numeric parameters (weights) rather than discrete symbolic rules. These models generally require learning from the data to adjust these parameters and improve their performance.

Since the late 20th century, sub-symbolic models, particularly neural networks, have dominated AI research due to their ability to automatically learn complex patterns from data. The typical workflow in training such models is as follows:

1. **Define the model architecture:** For example, choose a neural

network with a specified number of layers and connectivity.

2. **Choose a loss function:** This is a measure of error that the training algorithm attempts to minimise. For example, the loss function might quantify the difference between the model's predictions and the true labels of the training data.
3. **Train the model on the data:** Use an optimisation algorithm to adjust the parameters and minimise the loss on the training data. Common methods for this include gradient descent and its variants.
4. **Evaluate and iterate:** Test the model on separate data to ensure that it generalises well. Adjust the model or algorithm as necessary, which may include hyperparameter tuning and other refinements.

If the model learns effectively, it can then be deployed to make predictions or decisions on new unseen inputs.

A fundamental component of modern AI models, especially in deep learning, is the artificial neural network (ANN). Loosely inspired by the structure of brain neurones, ANNs consist of interconnected layers of simple units that transform input into output. Each connection has a weight, which is learnt during training. Figure 1 illustrates a simple example of a feedforward neural network with an input layer, a hidden layer, and an output layer. During training, data are passed through the network, producing predictions. The errors are then propagated backward through the network (a process called *backpropagation*) to adjust the weights [6]. This process allows the network to gradually learn representations of the data that are useful for specific tasks, such as classifying images or understanding language.

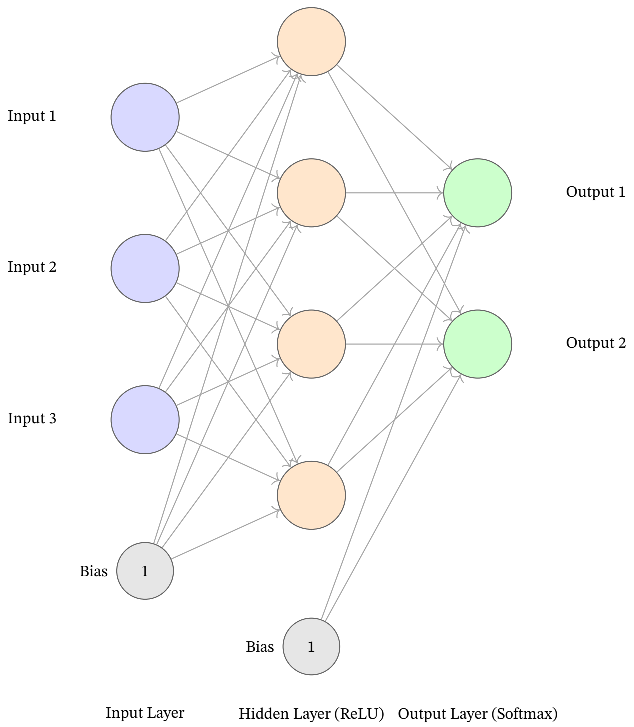

Figure 1. A compact feedforward artificial neural network for two-column layout. Includes three input features, a hidden ReLU-activated layer with four neurons and bias, and a Softmax output layer. All connections are fully connected and directional.

Neural networks are a flexible class of models: by increasing the number of neurones and layers (depth), they can approximate extremely complex functions. Deep networks automatically learn hierarchical feature representations. For example, in an image recognition network, early layers might learn to detect edges, midlayers compose edges into shapes, and later layers recognise objects. This automatic feature learning is a major advantage over previous AI approaches that required manual feature engineering.

It is important to note that not all AI is based on neural networks. Other models, such as decision trees, random forests, gradient boosting machines, support vector machines, and Bayesian models, are also widely used, particularly for tasks involving tabular data. However, in domains such as computer vision, speech recognition, and natural language processing (NLP), neural networks (especially deep learning models) currently achieve state-of-the-art results. As a result, these models form the focus of most technical discussions in AI.

3.2. Learning Paradigms: Supervised, Unsupervised, and Reinforcement Learning

The way an AI system learns can vary. The field of machine learning generally recognises three main paradigms of learning:

- **Supervised Learning:** The model is trained on input-output pairs, that is, labelled data where the desired correct output is provided for each input. The learning algorithm adjusts the model to best map inputs to outputs. This is analogous to learning with an answer key by a student. *Example:* Predicting house prices from features (size, location) using a dataset of past home sales (with price labels). The algorithm might be linear regression or a neural network, and it will learn to predict the price given the features by minimising the prediction error on the training set [3], [6]. Common supervised tasks include classification (the output is a category) and regression (output is a numeric value).
- **Unsupervised Learning:** The model is given data without explicit labels or targets, and it must find a structure in the data on its own. This is like discovering patterns or groupings inherent in the inputs. *Example:* Clustering of customers into segments based on purchasing behaviour, without being told any pre-defined categories. Algorithms such as clustering k-means will group data points that are similar in the feature space [9]. Other unsupervised tasks include dimensionality reduction (e.g., PCA), density estimation, and anomaly detection. Unsupervised learning is often used for exploratory data analysis or as a pre-training step.
- **Reinforcement Learning (RL):** The model (often called an *agent*) learns by interacting with an *environment*. Instead of direct labels, it receives feedback in the form of *rewards* for its actions [1], [5]. The goal is to learn a policy (a strategy that maps states to actions) that maximises the cumulative reward. This is akin to learning through trial-and-error experience guided by feedback. *Example:* A gamer agent who receives a +1 reward for winning or -1 for losing. During many simulated games, you will learn which actions lead to wins. Key concepts in RL include states, actions, rewards, and the notion of exploring the environment versus exploiting current knowledge [9]. Algorithms such as Q-learning or policy gradient methods are used to update the agent policy. Figure 2 illustrates the reinforcement learning loop: the agent observes the current state of the environment, takes an action, and in return gets a reward and the next state [6].

Each learning paradigm is suited to different types of problems. Supervised learning currently dominates industry applications because many tasks, such as object recognition, speech-to-text conversion, and predicting customer churn, can be framed with labelled datasets. Supervised learning tends to produce direct and high-accuracy solutions when there are ample labelled data available [3], [6]. Unsupervised learning is valuable when labelling is impractical as it can uncover hidden patterns or compress data. Reinforcement learning (RL) excels in scenarios that involve sequential decision making or where feedback is delayed, such as robotics, game play, or resource management. RL has achieved high-profile successes (e.g., AI in games,

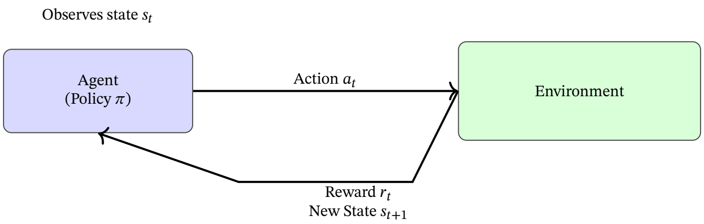

Figure 2. Reinforcement Learning loop: an agent observes a state from the environment, takes an action, and receives a reward and a new state. The agent's objective is to learn a policy for choosing actions that maximize the long-term reward[7].

energy optimisation of data centres), although it often requires many trial interactions, which can be a hurdle in real-world deployment [1], [5].

There are also hybrid approaches, such as semi-supervised learning (using a combination of labelled and unlabelled data) and self-supervised learning (where the data provide its own supervision, for example, by predicting part of the input from other parts). Recent large language models use self-supervised objectives, such as predicting the next word in a sentence, to learn from enormous unlabelled text corpora, effectively learning a wealth of knowledge without human annotation [9].

3.3. Key Concepts: Training, Generalisation and Model Evaluation

Regardless of the learning paradigm, several foundational concepts apply:

- **Training vs. Inference:** *Training* refers to the process of learning the parameters of the model from the data. This process is typically compute-intensive, especially for deep networks trained on large datasets, and is generally done offline. *Inference* is the application of a trained model to new data to make predictions or decisions. After the training phase (which can take hours, days, or more on specialised hardware), inference must be relatively fast to facilitate real-time applications.
- **Generalization:** The goal is to develop models that not only memorise the training data but also perform well on unseen data. The ability to generalise is crucial in AI. Techniques such as cross-validation, regularisation (e.g., weight decay, dropout in neural networks), and early stopping are used to prevent overfitting to training data. The performance of a model is typically assessed in a separate set of tests to estimate how well it generalises.
- **Evaluation metrics:** Depending on the task, different metrics are used to assess performance. For classification tasks, common metrics include accuracy, precision, recall, and the F1 score. For regression tasks, metrics such as mean squared error (MSE) are commonly used. In AI systems deployed in practice, additional metrics, such as fairness, robustness, and interpretability, can also be considered.
- **Optimization:** Training often boils down to an optimisation problem: minimising the loss function. Gradient-based optimisation methods, such as stochastic gradient descent (SGD) and Adam, are the workhorses for training neural networks. These algorithms iteratively adjust the model's weights in the direction that most reduces the error on a batch of training examples.
- **Hyperparameters:** These are parameters external to the model weights that affect the learning process (e.g., learning rate, network depth, and regularisation strength). Hyperparameters are typically tuned through experimentation. Automated hyperparameter tuning, using techniques like grid search, random search, or Bayesian optimisation, is common to find an optimal configuration.

To illustrate, consider a concrete example: training a neural network to classify images of handwritten digits (the MNIST dataset). We have 60,000 labelled examples, each being a 28×28 image with a known digit from 0 to 9. We choose a network architecture (e.g., 2 hidden layers with ReLU activation) and initialise weights randomly. In training, at each iteration, we:

1. Take a batch of images and perform a forward pass to get the predicted probabilities for each digit.
2. Compute the loss (e.g., cross-entropy between predicted probabilities and the true labels).
3. Calculate the loss gradients with respect to each weight using backpropagation.
4. Update the weights slightly in the opposite direction of the gradient (with a step size determined by the learning rate).

This process repeats over many epochs (passes through the data). Over time, the network predictions become more accurate in the training images. We monitor the accuracy of a validation set to ensure that the model is not over-fitting. Once training yields good validation performance, we evaluate the model on a test set and, if satisfactory, proceed to deploy it. This pipeline is typical for supervised learning problems.

In reinforcement learning, the loop differs (as shown in Figure 2), but the concept of iteratively improving a policy based on feedback (reward signals) is analogous to using gradients to improve a supervised model based on error signals. Both approaches are iterative improvement processes, guided by an objective.

3.4. AI System Components and Infrastructure

Beyond the learning algorithms and models themselves, practical AI systems involve substantial surrounding infrastructure.

- **Data pipelines:** 'Data is the lifeblood of AI.' Data preparation, including collection, cleaning, labelling, and augmentation, often consumes most effort. Large-scale AI applications require robust data pipelines and may also necessitate real-time data streams.
- **Computing hardware:** Training modern deep learning models often requires specialised hardware such as GPUs or TPUs (Tensor Processing Units). The rise of AI has gone hand in hand with advances in hardware, with the availability of GPUs around 2010 serving as a key enabler for the deep learning revolution [3], [6]. Today, specialised AI accelerators and cloud computing resources are widely used to support large-scale AI tasks.
- **Frameworks and libraries:** Tools such as TensorFlow, PyTorch, and scikit-learn provide high-level building blocks for implementing models and algorithms efficiently. These frameworks abstract much of the complexity of gradient computation and parallelisation, significantly accelerating AI development.
- **Deployment and integration:** AI models must be integrated into applications. This could involve converting a trained model to run on edge devices, which have constraints on memory and power, setting up API endpoints for a model serving service, or building user interfaces for AI functionality (such as a chatbot interface to a language model). In addition, monitoring model performance in production and setting up feedback loops for continuous learning are critical considerations.
- **Security and robustness:** Engineering AI systems includes securing them in terms of cybersecurity and ensuring their robustness to adversarial input. Adversarial attacks, where subtly modified inputs deceive a model, are an active area of research and a significant concern for practical AI deployment.

From a technical perspective, what distinguishes AI systems is the emphasis on learning and adaptation. Unlike traditional software, where every rule is hand coded by programmers, AI systems (especially those based on machine learning) *derive* their behaviour from data. This characteristic makes them powerful, but it also introduces new challenges, as their behaviour is implicitly defined by the data and training process, rather than explicit rules. We will explore the implications of this when discussing limitations such as bias and interpretability.

To conclude this section: Modern AI operates through complex models (often neural networks) trained on large datasets using iterative optimisation algorithms. These models can achieve impressive feats, such as vision recognition, language understanding, and strategic game play, by extracting patterns from data that humans would find difficult to express as explicit rules. With this technical foundation in place, we now turn our attention to how these AI capabilities are being applied across key sectors and the impact they are having.

4. Implications for IT, Finance, and Data Research

Artificial intelligence is not just a theoretical research domain; it has become a critical practical technology in various industries. In this section, we focus on three key areas: the IT sector, finance, and data-driven research to explore what AI means for modern society and how it is being leveraged in these domains.

4.1. Transforming the IT Sector

In the information technology (IT) industry, AI has driven automation and intelligent tools, reshaping how software and IT services are delivered. There are several clear ways AI is making an impact in IT:

- **Software Development and Quality Assurance:** AI techniques are increasingly used to assist in writing code (e.g., code completion, automated code reviews) and in testing software. For example, machine learning models can learn from large code repositories to suggest code snippets or identify common bugs. In quality assurance (QA), AI-based tools automate regression testing by intelligently generating test cases and detecting anomalies in software behaviour. This can significantly accelerate release cycles. AI's pattern recognition ability allows it to detect recurring error patterns in log files or code changes, enabling faster debugging [3], [6]. In general, by taking over repetitive and labour-intensive aspects of development and testing, AI allows human developers to focus more on design and creative aspects.
- **IT Operations and Infrastructure (AIOps):** Managing complex IT systems (data centres, cloud infrastructure, corporate networks) generates vast amounts of log and performance data. AI helps by sifting through this data to detect incidents, predict outages, and optimise resources – a field commonly referred to as 'AIOps' (AI for IT Operations) [9]. For example, AI algorithms can predict when a server is likely to fail based on sensor data and logs, enabling pre-emptive maintenance. They can also automate responses to common incidents (self-healing systems). Gartner coined the term AIOps to describe multi-layered platforms that use big data, analytics, and machine learning to automate IT operations processes such as monitoring, service desk management, and automation [1]. The result is improved uptime and efficiency in IT environments that have become too complex for manual human monitoring alone.
- **Service Management and Support:** AI is also applied to IT service management, such as handling user requests, helpdesk tickets, and other support functions. AI-powered chatbots and virtual assistants are deployed to manage routine helpdesk queries (e.g., password resets or frequently asked questions), reducing the workload on human support staff. Natural language processing (NLP) enables these systems to understand user questions and either answer them or route them to the appropriate solution. AI can also help prioritise and classify IT tickets by analysing their content (e.g., urgent issues can be auto-flagged). By applying AI to service management, organisations improve response times, improve user satisfaction, and reduce support costs [5], [6].
- **Security and Threat Detection:** Cybersecurity is a critical area of IT, and AI has become an indispensable tool to detect and respond to threats. Machine learning models can identify patterns of normal versus malicious behaviour in network traffic, user logins, or application usage. This enables the real-time detection of intrusions, fraud, or abuse. For example, AI-based security systems might analyse millions of log-ins to detect anomalies indicative of credential-cramming attacks or monitor network packets to flag patterns matching known malware signatures. The dynamic nature of cyber threats makes the ability of AI to continuously learn and adapt especially valuable. Many companies now rely on AI-driven Security Information and Event Management (SIEM) systems for automated threat detection [3].

These AI-driven changes in IT have broad implications. The efficiency and reliability of IT services are improved: systems experience fewer outages due to predictive maintenance, and issues are resolved faster with intelligent support. At the same time, the role of IT professionals is evolving – they are increasingly supervising AI tools or focussing on higher-level strategy, rather than performing all tasks manually. In general, AI serves as a force multiplier for IT operations.

Industry surveys show that the adoption of AI is becoming essential for the delivery of competitive IT services. For example, one study predicts that global spending on AI systems will exceed $500 billion by 2027 in areas including IT operations and business process management [8], [9]. Many IT firms now promote AI integration as part of their offerings (often referred to as 'AI driven' services or 'self-driving IT'). AI in the IT sector is about automation, intelligence, and the ability to manage complexity at scale. Companies that successfully leverage AI in their IT processes often achieve more agile and resilient operations compared to those relying purely on manual human effort.

4.2. AI in Finance: Automation, Analytics, and Decision Support

The financial services industry was an early adopter of AI technologies, and today AI's influence in finance is pervasive and continues to grow. Major financial institutions invest heavily in AI to improve productivity, decision-making, and customer experience [8], [9]. Some key applications and implications of AI in finance include:

- **Algorithmic Trading and Investment:** AI-driven algorithms now execute a significant share of trades in stock, forex, and other markets. These systems can analyse market data at superhuman speeds and make rapid trading decisions based on patterns or signals (sometimes in fractions of a second). Machine learning models also help in portfolio management by predicting asset price movements or optimising asset allocations. Hedge funds and investment firms use artificial intelligence models to design complex strategies that automatically adapt to market conditions. The benefit is more efficient markets and the ability to exploit very subtle inefficiencies; however, it also raises concerns about flash crashes and the need for oversight, as AI agents might collectively behave in unforeseen ways.
- **Risk Management and Fraud Detection:** Risk management is central to finance, and AI provides more accurate and granular risk models. For example, banks use machine learning to improve credit scoring by analysing the data of a borrower in much more detail than traditional linear models, allowing better predictions of default risk [6]. AI models can incorporate alternative data (such as online footprints or transaction history patterns) to make credit decisions for people without extensive credit histories, potentially expanding financial inclusion. In trading, AI is used for real-time risk monitoring (e.g., calculating Value-at-Risk with models that adapt to current volatility regimes). **Fraud detection** is another critical area: credit card companies and payment processors employ AI to detect fraudulent transactions by recognising anomalies in spending behaviour. These

models operate on streaming data, flagging suspicious activity (e.g., an unlikely location or high-value purchase) within milliseconds, often preventing fraud or limiting losses. By analysing vast amounts of transactional data, AI systems have achieved high accuracy in catching fraud while reducing false alarms, thus protecting consumers and institutions.

- **Automation of Processes (Robotic Process Automation - RPA):** Finance has many repetitive, document-driven processes (e.g. mortgage approvals, insurance claim processing, accounting reconciliations). Artificial intelligence, often in conjunction with RPA tools, is automating many of these back-office operations. For example, AI-powered document analysis can read and interpret forms, invoices, or contracts (using computer vision and NLP), extracting relevant information for further processing. A concrete case is an insurance company that uses artificial intelligence to automatically approve straightforward claims by analysing submitted documentation and cross-referencing the details of the policy. Automation reduces processing time from days to minutes and minimises human error. One reported effect is that employees can be redeployed from routine paperwork to more value-added or analytical tasks. Financial firms see significant cost savings and faster service delivery through these AI-driven efficiencies.
- **Customer Service and Personalization:** Just as in IT support, financial institutions use AI-powered chatbots to handle customer inquiries (e.g. balance requests, simple FAQs) through chat or voice, available 24/7. AI can also personalise financial advice: so-called ’robo-advisors’ provide automated investment advice tailored to the individual’s needs.

Leaders in finance view AI as essential: A Columbia Business School report notes that in 2023, financial services companies spent about $35 billion on AI, with that number expected to nearly triple to $97 billion by 2027 – the fastest growth of any major industry [9]. The arms race in AI is driven by the competitive edge it provides: better predictions directly translate to profit in trading; better risk models mean lower losses and capital savings; better service results in happier customers and higher retention.

At the same time, the adoption of AI in finance is urging regulators to update frameworks. Issues such as algorithmic transparency, fairness in credit decisions, and the systemic risks of AI-driven markets are hot topics. Financial regulators are working to ensure that, as banks rely on AI, they still manage to explain decisions (for example, why someone was denied a loan) and maintain accountability. In general, AI in finance increases human decision making with data-driven insights and efficiency, but does not eliminate the need for human oversight. As one panel of industry experts put it, AI is ’a tool, like a screwdriver’ that can greatly enhance capabilities, but humans remain crucial for providing judgment and ensuring that AI’s output is applied correctly [1], [8].

4.3. AI in Data-Driven Research and Science

Beyond specific industries, AI has become a transformative tool in scientific research and any field that relies on extracting knowledge from large datasets (often referred to as "data science"). Researchers are increasingly using AI to handle data volumes and complexities that human analysis could never manage alone. Here are a few notable examples:

- **Healthcare and Biomedical Research:** AI is accelerating drug discovery and genomics. For instance, deep learning models can screen billions of chemical compounds to predict which might have therapeutic effects on a target protein, dramatically narrowing down candidates for lab testing. In genomics, machine learning helps identify patterns in DNA that correlate with diseases, or predict the 3D structure of proteins (as demonstrated by DeepMind’s AlphaFold solving protein folding, which can aid in understanding diseases and developing drugs) [9]. AI is also used in medical imaging: radiologists now have AI assistants that can detect tumors or lesions in X-rays, MRIs, etc., sometimes earlier or with equal accuracy to human experts [6]. The implication is faster diagnoses and potentially new cures discovered more quickly. Of course, these AI systems undergo rigorous validation since lives are at stake, and they typically assist rather than replace medical professionals.
- **Scientific Research (Physics, Astronomy, Climate Science, etc.):** Many scientific domains have massive data streams – telescopes surveying the sky, particle colliders generating collision data, sensors monitoring the climate. AI is indispensable in analyzing this data. In astronomy, AI models classify astronomical objects (e.g., stars, galaxies, supernovae) in sky survey images and have even been used to discover new exoplanets by sifting through satellite data for the faint signatures of distant planets [3]. In physics, AI helps identify rare events in particle collision data that might indicate new particles or phenomena. Climate scientists use AI to improve models for weather prediction and climate projections by learning complex patterns from historical data. What all these fields have in common is that AI augments human researchers’ ability to make sense of *Big Data*, often revealing subtle patterns or correlations that a human might miss. AI can also act as a "multiplicative" factor – for example, an AI-driven simulation might allow exploring thousands of climate policy scenarios quickly to predict potential outcomes.
- **Data Analysis and Knowledge Discovery:** Even in fields like social sciences or humanities, where data may be text, audio, or video, AI is unlocking new kinds of analysis. Natural language processing (NLP) can analyze millions of documents or social media posts to identify trends in public opinion or trace the evolution of ideas. An area known as "digital humanities" uses AI to, for example, analyze literary texts for themes or patterns on a scale previously impossible. In economics, AI models analyze financial news and reports to quantify sentiment or predict economic indicators. In all these cases, AI acts as an intelligent assistant that can quickly summarize or find structures in overwhelming amounts of information, which researchers can then interpret. One often-cited benefit is that AI can generate hypotheses by finding unexpected patterns, which human experts can then investigate further – effectively providing a new way to derive insights from raw data.

A notable observation is that AI itself has become a subject of research not only in computer science but across disciplines interested in intelligence, cognition, and complex systems. For example, cognitive scientists collaborate with AI researchers to use neural networks as models to understand human cognition (comparing how an AI vision model and a human brain respond to the same images, for example). In economics and ethics, the rise of AI prompts research into its societal effects.

In terms of implications for society, AI in research accelerates the pace of discovery. Previously intractable problems (such as analysing trillions of possible protein folds or simulating quantum chemistry accurately) are now seeing breakthroughs [3], [9]. This means potential benefits such as faster development of medical treatments, more accurate projections of climate change to inform policy, and a deeper understanding of fundamental science. It also means that the skill set for researchers is shifting: computational literacy and the ability to leverage AI tools are becoming essential even in fields that were traditionally more theoretical or experimental. We are witnessing a paradigm shift where discovery is increasingly *data-driven*, with AI acting as the engine that sifts through the data [1], [6].

In the domains of data research, AI serves as a powerful amplifier of human analytical capabilities. It can be considered “the great equaliser of big data,” turning the flood of data we now collect into actionable insights [9]. This has democratising effects too: An individual scientist or a small startup can now harness cloud AI services to analyse big data without needing a supercomputer of their own, lowering barriers to entry for innovation.

Across IT, finance, and research, a common theme emerges: AI systems excel at absorbing large amounts of data and identifying optimal patterns or actions, leading to greater efficiency, personalization, and even new capabilities (such as forecasting that was previously unattainable). However, deploying AI on scale also raises important considerations, some of which we will discuss in the next sections on what AI is less capable of and the risks involved.

5. Strengths of Contemporary AI

Having explored where AI is applied, we now distill the types of tasks for which current AI techniques (especially machine learning and deep learning) are particularly well suited. The strengths of AI align with certain characteristics of problems, and recognising these helps in deciding when to employ AI solutions. In the following, we discuss several categories of tasks in which AI tends to perform exceptionally well, with examples from IT, finance, and data science as illustrations.

5.1. Pattern Recognition and Perception

One of AI's greatest strengths is identifying patterns in large and complex datasets, in many cases exceeding human ability in terms of precision, scale, or consistency. Deep learning, in particular, has revolutionised perceptual tasks:

- In computer vision, AI models can recognise objects, faces, and scenes in images or videos with high accuracy. They can detect tumours on medical scans, read handwritten text, or monitor products on an assembly line for defects. Such visual pattern recognition tasks, which involve subtle variations and high-dimensional pixel data, play to the strengths of convolutional neural networks. For example, an AI vision system can be trained on millions of images to distinguish hundreds of cat breeds, a level of fine-grained differentiation that a human might struggle with. Consistency is also a key factor: unlike humans, AI is not fatigued or less attentive after reviewing thousands of images.
- In audio and speech, AI models (such as recurrent or transformer networks) can recognise speech to transcribe it, identify speakers, or even detect emotion from tone. Tasks like converting speech to text (as used in virtual assistants or automated captioning) are now highly accurate. AI has also been used to detect patterns in sounds, such as identifying mechanical faults from engine noise or detecting calls of endangered animal species in audio recordings. Again, these are pattern recognition problems where AI can sift through more data than a human could and pick out telltale features.
- In data science broadly, pattern recognition manifests itself as finding correlations and clusters in data. For instance, in finance, AI can detect patterns in transaction data that indicate fraud (as previously mentioned) or uncover nonobvious relations between market indicators that support investment decisions. In IT operations, anomaly detection (for example, spotting patterns in system metrics that precede a failure) is a pattern recognition task in which AI excels [9].

The underlying reason for AI prowess in these tasks is its ability to handle high-dimensional inputs and complex nonlinear relationships. Neural networks, with sufficient data, can approximate extremely complicated functions, which is necessary for tasks like image or speech recognition. Moreover, AI can integrate information from multiple sources (for example, a surveillance system that combines video and audio patterns). The key limitation - which we will revisit - is that this works best when ample labelled data is available to learn from, particularly in supervised settings.

5.2. Processing Big Data Quickly and Efficiently

AI algorithms can sort and analyse vast amounts of data far faster than humans. This makes AI indispensable for big data analytics. Tasks that involve scanning millions of records to find insights or scanning network logs for intrusion attempts are well-suited to AI because:

- The volume of data is too large for manual analysis, but AI thrives on volume. In fact, more data often improves the accuracy of an AI model. For example, a recommendation system on an e-commerce platform may analyse browsing and purchase data from hundreds of millions of user interactions to find patterns such as 'people who buy X also like Y'. Doing this without AI would be impossible, but a machine learning model (such as matrix factorisation or a deep learning recommender) can handle it and continuously update as new data come in.
- AI can handle the velocity of data, i.e., streaming data analysis. In finance, high-frequency trading algorithms can process market tick data in microseconds and execute trades accordingly. In IT monitoring, an AI system might continuously ingest log streams and metrics, issuing an alert the moment an anomaly is detected. Humans cannot match this real-time processing speed when faced with such data firehoses [6].
- Efficiency and scalability: With appropriate hardware, AI computations can be parallelised. GPUs can perform thousands of operations in parallel, allowing AI models to process data quickly. This is why tasks like training a deep network on ImageNet (more than 14 million images) or translating an entire Wikipedia-worth of text are feasible with AI. From a business standpoint, tasks that previously required large analysis teams to combing data can now be done with an AI system and a handful of analysts to interpret results, significantly reducing cost and time.

An example of data research: AI models have been used to analyse the equivalent of 100,000 years of climate simulation data in a short time, identifying patterns of extreme weather events [9]. This demonstrates how AI does not get overwhelmed by the data scale. In fact, AI often finds subtle signals precisely using the scale (e.g., identifying a 0. 1% occurrence pattern that is only noticeable when you have millions of samples).

5.3. Automation of Repetitive and Structured Tasks

AI is well-suited for tasks that are routine, repetitive, or structured – especially those that involve making numerous small decisions or classifications based on data. Some examples include:

- **Data entry and processing:** AI-powered optical character recognition (OCR) can read documents and digitise them, a task that once required an army of typists. Now, entire archives can be digitised with minimal human intervention. Similarly, AI can reconcile invoices, match records, or flag inconsistencies in large datasets – tasks traditionally performed by clerks.
- **Manufacturing and Robotics:** In manufacturing settings, robots equipped with AI vision can pick and sort objects, assemble components, or inspect products for quality at high speed. These tasks are repetitive (involving the same movements or checks on thousands of parts) and structured (with a controlled environment), making them ideal for AI automation. The result is often higher throughput and precision. For example, an AI robot might place components with sub-millimeter accuracy consistently, reducing error rates.
- **Customer interaction automation:** AI chatbots handle repetitive customer queries ("What's my balance?", "When will my order arrive?") multiple times a day. AI can answer these queries consistently and instantaneously, functioning like a tireless employee. Many companies report that AI chatbots resolve a large

percentage of customer queries without needing human escalation, thus dramatically scaling their support capacity.
- **Email filtering and routing:** A common example is spam filters in email. AI filters out junk mail with high accuracy by learning patterns of spam. Similarly, AI systems in enterprises automatically categorise and route emails or support tickets to the appropriate department (sales inquiry, technical support, billing issue, etc.) by analysing the content. These mundane tasks are handled invisibly by AI in the background to streamline workflow.

The advantage of AI in these tasks is not only labour saving, but also consistency and speed it brings. AIs do not get bored or tired, so repetitive tasks performed by AI will have a low error variance. As noted in the context of the AEC (architecture, engineering, construction) industries, AI significantly boosts productivity by taking on mundane tasks, for example, generating routine design documentation or performing cost estimations, allowing professionals to focus on creative or complex aspects [6], [9]. In an HR example, AI might screen resumes to shortlist candidates based on set criteria, doing what could take a recruiter many hours.

5.4. Prediction and Forecasting

AI systems, particularly machine learning models, excel at making predictions from historical data. If a task can be framed as 'given these inputs, predict that output', and there are many examples to learn from, supervised learning often produces a model that outperforms traditional statistical methods. Some examples include:

- **Predictive maintenance:** In industry, AI models predict when machines or components are likely to fail by learning from sensor data patterns that preceded past failures. This allows companies to replace or service equipment *before* if a failure occurs, minimising downtime. For example, an AI might monitor vibration and temperature readings from a turbine and predict a failure two weeks in advance with high confidence, scheduling maintenance proactively [6].
- **Demand forecasting:** Retailers use AI to forecast product demand at a granular level. Traditional forecasting might look at monthly sales; AI can predict daily or even hourly demand for each store and each product by considering factors like promotions, weather, local events, etc. This fine-grained forecast optimises inventory levels, ensuring that shelves are stocked but not overstocked. Amazon, for example, uses AI for 'anticipatory shipping' – predicting what you will order and moving it to a nearby warehouse even before you order, based on patterns.
- **Financial forecasting:** AI is used to predict stock prices, market trends, credit defaults, and macroeconomic indicators. Models can incorporate a multitude of signals – technical indicators, sentiment from news, historical correlations – to make short-term or long-term forecasts. Although not infallible, these models often capture complex relationships that simpler models miss, giving financial firms an edge. For example, an AI might predict intraday price movements for dozens of stocks simultaneously, helping a trading desk position itself advantageously [3].
- **Personalization (predicting user preferences):** When Netflix or YouTube recommends content, or a music app creates a 'discover weekly' playlist for you, it essentially predicts what you will like based on past behaviour. AI recommendation engines predict the rating or click-through probability for each user-item pair and then present the top predictions. This predictive ability to tailor experiences is a major strength of AI. It has a wide usage across domains from e-commerce (predicting which product a user is likely to buy next) to online advertising (predicting who will click on an ad) [9].

What sets AI prediction apart is its ability to model very complex, non-linear interactions in the data. Traditional forecasting might rely on linear regression or time series models like ARIMA, which have limitations. AI can ingest many more variables and find hidden nonlinear effects (e.g., how the combination of weather, day of week, and a specific promotion drives sales in a particular store). Furthermore, AI can update predictions in real-time as new data comes in, which is critical in dynamic environments.

However, it is important to monitor such models because if conditions change (e.g., a pandemic radically changes consumer behaviour), the patterns learnt from history may no longer hold - something we will touch on when discussing limitations like generalisation.

Task Type AI Excels At	Examples and Domains
Pattern Recognition	Image classification in diagnostics, speech recognition (e.g., virtual assistants), and anomaly detection in network traffic.
Large-Scale Data Analysis	Real-time fraud detection on large financial datasets, customer segmentation in marketing, and scientific data mining in astronomy and genomics.
Repetitive Task Automation	Quality inspection in manufacturing, chatbots for routine queries, and automated data entry via OCR in enterprise settings.
Predictive Modeling	Predictive maintenance in utilities, personalized content recommendation systems, and demand forecasting in retail logistics.
Complex Decision-Making and Optimisation	Strategic game AI (e.g., Go, StarCraft), portfolio optimisation in finance, and routing optimisation for delivery logistics.

Table 2. Illustrative examples of domains where AI systems are particularly effective.

5.5. Handling Complexity and Multivariate Relationships

Finally, AI excels in domains that are too complex for explicit human reasoning. In many cases, there are problems without an analytic solution or an easy rule-based approach, but AI can approximate a solution by brute-force learning.

An example is in 'playing the game': For games like Go or chess, the number of possible states is astronomical. Traditional algorithms struggled with Go until deep reinforcement learning emerged to approximate the value of positions and policies through self-play. The success of AlphaGo highlighted how AI can handle immense combinatorial complexity, discovering strategies that even human champions had not considered [9]. Similarly, in multiplayer video games or complex simulations, AI agents learn to navigate environments that have enormous state spaces and interacting factors.

In engineering design and optimisation, AI techniques can tackle multiobjective optimisation problems. For example, designing an aircraft component involves trade-offs between weight, strength, aerodynamics, cost, and more. AI (including techniques such as genetic algorithms or neural networks) can search this complex design space to propose solutions that meet all criteria, some of which a human designer might not have conceived. The phrase 'AI can explore numerous design possibilities much faster and more extensively than before' has been observed in the context of AEC (architecture, engineering, construction) [3], [6].

Another area is multivariate analytics – where outcomes depend on many interdependent variables. For example, in medicine, predicting disease progression might depend on genetic factors, lifestyle, environment, etc., in highly nonlinear ways. Artificial intelligence models (such as deep networks) can integrate these multivariate relationships. They might identify that a combination of subtle readings in blood tests, when seen together, is predictive of a certain condition -something that no single medical indicator reveals on its own.

To illustrate in a data science context, consider trying to model customer churn for a subscription service. Churn might depend on dozens of features, such as usage frequency, customer service interactions, demographics, and competitor presence, with complex interactions (e.g., high usage might usually indicate loyalty, but if accompanied by repeated service complaints, it might predict churn). An AI model can learn this intricate interplay automatically, whereas a manual analysis might miss such second-order combinations.

Tasks characterised by high-dimensional data, complex rules, or massive possibilities, where writing a fixed programme would be impractical, are fertile ground for AI. These are precisely the scenarios where AI's ability to learn and adapt gives it an advantage. Table 2 summarises some of the tasks and examples of AI's strengths.

5.6. Lack of Contextual Understanding and Common Sense

Perhaps the most notorious weakness of AI is its lack of genuine understanding. AI models don't possess common sense - the basic level of practical knowledge about the world that humans take for granted. They also do not truly grasp context or meaning; they operate on surface correlations in data. This leads to a variety of issues.

- **Misinterpretation of language or vision without context:** Natural language processing models might interpret a sentence literally and miss the implied meaning, sarcasm, or cultural references that a human would catch. For example, an AI assistant might interpret 'Can you tell me how to get out of a speeding ticket?' as a factual query about legal procedure, while a human might recognise it as someone looking for unethical advice (or a joke). AI lacks the situational awareness to navigate such nuance. Similarly, in computer vision, an AI might correctly label objects in an image but may not understand the situation depicted: it may see people running and classify it as'sport' when, in fact, they are fleeing danger.
- **Failure at reasoning tasks that require understanding of concepts:** AI is famously struggling with seemingly simple common sense questions. For example, a classic example: if I put my socks in a drawer and close it, then I open the drawer later, are the socks still there? A human knows that the socks will still be there; an AI language model might get this right, but not because it 'knows', rather than because it has seen similar statements during training. If posed differently ("I put my socks in the drawer, went away for a week, and no one touched the drawer; Where are my socks?"), some AIs might still get the answer wrong. A large-scale test known as the 'Winograd Schema Challenge' (a common-sense-requiring pronoun understanding test) has been tough for AI. Although there is progress with enormous language models, they still make mistakes that reveal a shallow grasp of meaning.
- **Rigidity and literalism:** Because AI does not have a true understanding, it cannot easily adapt instructions to intent. If you slightly misinterpret an input, an AI might fail where a human would infer your intent. For example, an AI home assistant may not turn off the lights when you say 'I am going to bed now' because it was not explicitly told as a command, whereas a human butler would get the hint. An example from an educational blog notes that AI responses can feel 'robotic and impersonal, lacking depth of human interaction,' precisely because they do not grasp the emotional or contextual subtext [9].
- **No true understanding of causality:** AI often confuses correlation with causality. You may notice that in training data, when the grass is wet, it usually rains. But if you then water the lawn with a hose (wet grass without rain), a naive AI weather system might erroneously predict rain. Humans understand causes; AI largely does not, unless explicitly trained in causal inference (which is an active research area, but not solved). This ties into issues like susceptibility to spurious correlations: an AI might predict that a customer will default on a loan because they use all caps in their emails (perhaps that correlated in historical data for some odd reason), which is clearly not a causal factor but could slip into a model if not carefully controlled.

The root of these issues is that current AI lacks a model of the world. It does not know physics, social norms, or basic facts unless those were implicitly encoded in the training data and model weights. There is an infamous quote: "AI is only as good as its data"; If the data do not cover some scenario or contain some knowledge, the AI is unaware of it. Humans, on the other hand, have broad common sense knowledge. For instance, we know that objects fall down, not up, that people have motivations, that time has an order, etc. AI inherently does not know all that.

This can lead to dangerous mistakes. Consider an AI vehicle that does not grasp context: a pedestrian waving might be interpreted by vision as simply 'person' without recognising that waving means 'go ahead' or 'thank you'. Or an AI content filter might ban a post discussing violence in a historical context because it sees violent words without the context that it is educational.

Researchers attempt to inject common sense into AI by building knowledge graphs or training on massive datasets (the hope is that AI will implicitly absorb some common sense). Large language models have in fact learnt a lot of factual knowledge (it's stunning that they can answer trivia). However, they still lack a deeper understanding. One blog on AI limitations succinctly stated: "AI systems struggle with context understanding and lack common sense reasoning... limiting their ability to interpret complex human language and emotional nuances" [9]. The consequence is that for any situation requiring flexible, context-aware reasoning or creativity, humans still have a definitive edge over AI.

5.7. Data Dependency and Bias

AI's capabilities fundamentally hinge on data – "garbage in, garbage out" remains a pertinent adage. There are several facets to this:

- **Data Hunger:** Most AI models require large amounts of data to train effectively. In domains where data are scarce or expensive to obtain, AI models may perform poorly. For example, developing an AI diagnostic for a very rare disease is challenging because there are not enough case examples to learn from. In contrast, humans can sometimes generalise from just a few examples by applying prior knowledge. AI often lacks that, unless transfer learning is feasible from a related domain.
- **Sensitivity to Data Quality:** If the training data contain errors, noise, or inconsistencies, the AI will learn from those as well. A model might latch onto random fluctuations as if they were meaningful (overfitting), leading to poor performance on new data. Also, if the data collection process changes over time (e.g., a different sensor is used, or a survey question is re-worded), the model might start failing unless re-trained. Essentially, AI is as good as the signal in the data; it cannot magically overcome fundamentally bad data.
- **Bias in Data leading to Biased AI:** AI systems notoriously inherit biases present in their training data [9]. If historical data reflect human biases or systemic biases, AI will often reproduce or even amplify them. For example, a hiring algorithm trained on past hiring decisions at a company might learn to discriminate against candidates from a group that was historically underhired, not because that trait impacts job performance, but because of bias in the historical decisions. There have been multiple high-profile cases: facial recognition systems less accurate on darker-skinned individuals because the training set was skewed towards lighter-skinned faces; or language models generating stereotypical or derogatory text about certain groups because of biased text in their training corpus. This is a serious limitation because it can lead to unfair or unethical outcomes if artificial intelligence is used in decision making (e.g., credit, employment, and police). Ensuring fairness requires careful data curation and algorithmic bias mitigation, which is an active area of research and policy [3], [6].
- **Lack of Adaptability to Data Shifts without Retraining:** If the world represented by the data changes (known as 'concept drift'), AI models typically will not adjust on their own unless

they are explicitly retrained with new data. For example, an AI trained to predict consumer preferences in 2019 might have been thrown off by the radically different patterns during the 2020 pandemic. Humans can often adapt quickly by recognising the change in context, but an AI could continue making predictions as if nothing changed, yielding poor results. Regular retraining and model monitoring are required to keep AI systems relevant, which is an overhead that is sometimes not fully appreciated.

These issues underscore that AI is not one-and-done software that you can set and forget; it is part of a data ecosystem. The phrase 'data dependency' also implies that AI performance is upper bound by the content of the information in the data. If some critical factor is not captured in the data, the AI cannot learn it. For example, if medical records lack a certain symptom because doctors did not record it, an AI predicting diagnosis might completely miss the relevance of that symptom.

Data problems also cause AI to sometimes make obvious mistakes to humans. For example, an image classifier might label a picture of a panda as a gibbon because some statistical quirk misled it – something a human would almost never do because we intrinsically know what a panda looks like. An analysis of such errors often reveals an odd pattern in the training data or a heavy reliance on a background detail rather than the object shape (e.g., the panda was misclassified because of foliage that looked like the background in many gibbon photos). AI does not have the innate category concept, just statistical associations.

AI's dependence on data is a double-edged sword: it is the source of its power (learning from data rather than requiring explicit programming), but also a source of vulnerability - to bias, errors and context changes. Identifying data issues is arguably the most important part of any AI project. As one resource put it, ensuring diverse and representative data and robust training is key to mitigate unfair or erroneous outcomes [9].

5.8. Opacity and Lack of Interpretability

Most high-performing AI models today, especially deep neural networks, are often described as black boxes. That is, they can be extremely hard to interpret: we feed in an input, get an output, but have little insight into how or why the model produced that output. This lack of transparency or interpretability is a significant limitation in the domains where understanding the decision process is important.

There are several reasons interpretability matters:

- **Trust and Verification:** In sensitive applications such as healthcare or criminal justice, we cannot take the word of an AI without justification. A doctor needs to know *why* an AI recommended a certain diagnosis to trust it and act on it (did it find a pattern on the MRI that correlates with a disease, or did it latch onto an artefact in the image?). If an AI judge-assistant tool predicts that a defendant is a high flight risk, the judge must understand the reasoning. The opacity of AI currently makes it difficult to fully trust, as it can base decisions on spurious correlations or biases that we would not accept if we knew. One source notes that when people don't understand how an AI makes decisions, they are reluctant to use it [9]. In fact, the lack of interpretability causes a 'black-box effect', which can erode confidence and hinder adoption.
- **Debugging and Error Analysis:** When an AI system makes a mistake, it is often non-trivial to figure out exactly what went wrong internally. If a neural network misclassifies an image, we cannot simply inspect a few weight values and immediately see the error. We might have to resort to techniques like saliency maps (to see which pixels influenced the decision) or analyse the training data influences. This is an active area of research (Explainable AI, or XAI), which aims to provide explanations for model behaviour [3]. Until we have better interpretability tools, there is a risk that AI systems harbour hidden failure modes that we only discover in operation.
- **Accountability and Ethics:** If an AI system causes harm (e.g., a self-driving car has an accident or an AI incorrectly denied someone a loan), who is responsible? Part of that question is related to being able to explain what the AI did and whether it followed acceptable rules. Currently, many AI decisions are not easily backtrackable in human terms, which complicates accountability. It also challenges regulations such as the EU GDPR, suggesting that individuals have a right to an explanation for decisions made about them by algorithms.
- **Overreliance Risk ("Automation Bias"):** Paradoxically, the better AI is, the more people could overrely on it without question. If a system is usually right, humans can start rubber stamping its decisions, even when it is wrong (this is known as automation bias). Studies have shown, for example, that physicians assisted by an AI diagnostic might ignore contrary clinical evidence if the AI gives a certain result, especially if they cannot pinpoint why the AI could be wrong [9]. Overreliance can be mitigated if AI provides understandable reasons or uncertainty estimates, but with black-box models, users might either distrust them too much or trust them too much, both problematic. The Stanford HAI article on AI overreliance indicates that explainability is being looked at as a solution to prevent people from blindly trusting the output of an AI [9].

The black-box nature is particularly acute with deep learning. A deep neural net might have millions of parameters that form a complex nonlinear function; trying to directly understand how it makes decisions is extremely difficult; it is essentially a high-dimensional mathematical transformation without semantic annotations for each part. This is unlike a decision tree or linear model, where one could trace a path or weight to see the influence. Many are calling for 'Explainable AI', where models are inherently interpretable or come with tools that explain their reasoning in human terms [3].

There has been progress: techniques such as LIME or SHAP approximate features that strongly influenced a particular decision; deep networks can be probed to see what internal neurones respond to (e.g., one might respond to 'this looks like a human face' and another to 'this looks like text' within an image, giving some insight). But these are imperfect and sometimes themselves hard to interpret.

The bottom line is that current mainstream AI often lacks a clear explanation for its outputs. This is often summarised as the "black box problem" and is considered one of the main challenges to wider adoption in fields such as healthcare and finance [9]. As noted in a resource from PSMJ, the complex AI decision making process is not easily interpretable for humans, leading to a delay in use [9]. Another quip is that deep learning models are like a super-intelligent student who aces the test but you have no idea how they arrived at the answers.

This limitation reinforces the notion that for critical decisions, AI should augment rather than replace human judgment until we can verify and explain what it is doing. It's an area where sometimes simpler or more interpretable models are chosen over an inscrutable complex model, sacrificing a bit of accuracy for transparency – especially in regulated industries (this trade-off is an ongoing discussion in the AI community and among regulators).

5.9. Creativity, Emotions, and Security

There are a few additional areas where AI is commonly said to be weak:

- **True creativity and intuition:** While AI can generate art, music, or text that appears creative (e.g., procedural game content or AI-generated paintings), it doesn't have creativity in the human sense of intentional novelty or emotional depth. AI generation is based on recombination of patterns from training data, not genuine inspiration or understanding of aesthetic value. Thus, AI might produce a hundred variations of a melody, but choosing one that evokes a particular feeling or fits a cultural context is still a human strength. Similarly, AI-written prose might be grammatically perfect but often lacks the coherent intentional narrative that a human author provides. An AI could produce a "remix" of Shakespeare-style text, but it isn't going to invent a wholly new literary genre with purpose. As noted in a PSMJ resource, AI struggles with tasks requiring true creativity and intuition [9].

- **Emotional intelligence and empathy:** AI does not have emotions or empathy. It can simulate empathetic responses to some degree (e.g., a chatbot can be programmed to say "I'm sorry to hear that, that must be difficult"), but it doesn't genuinely understand or share feelings. In domains like mental health counseling or even customer service, this is a limitation – AI can offer facts or basic supportive phrases, but it cannot truly comfort or build rapport in the way a person can. This can make AI interactions feel unsatisfying or even inappropriate in sensitive situations. As PSMJ pointed out, current AI has "zero emotional intelligence" [6]. It cannot gauge a person's mood beyond maybe analyzing tone of voice or facial expression, and even then, it doesn't *feel* anything about it. This is why roles requiring human connection (like therapy, or negotiations) remain largely human.
- **Security and adversarial robustness:** AI models, especially neural networks, have a peculiar vulnerability: they can be fooled by adversarial examples. These are inputs that have been subtly modified to mislead the AI while appearing almost normal to a human. A classic example: adding an almost imperceptible noise pattern to an image of a stop sign can make an AI classifier see it as a speed limit sign, whereas any human still sees a stop sign [9]. This is a serious concern for security – imagine malicious actors causing an AI system to misclassify a critical input (e.g., making a biometric security system mistake one person for another via a specially crafted accessory, or causing an autonomous car to mis-read traffic signs using carefully placed stickers). AI tends to rely on all sorts of minute cues in data; adversaries can exploit that since the AI has no common sense to say "that's clearly still a stop sign despite the sticker." Ensuring AI is robust to such perturbations is hard. Additionally, AI systems could be attacked by feeding them harmful data during training (data poisoning attacks) which inject biases or backdoors. The bottom line is AI opens new attack surfaces, and the technology to secure and harden AI models is still maturing [3].
- **Resource intensity and environmental cost:** Training large AI models, especially deep learning models with billions of parameters, is extremely computationally intensive. This has practical and environmental downsides. Practically, not every organization can afford the hardware or cloud compute to train or even deploy these models (creating a bit of an AI divide). Environmentally, the energy consumption is a concern – some estimates claim that training a single big transformer model can emit as much carbon as five cars in their lifetimes. While this is more about current implementation than a fundamental inability, it's a limitation in the sense that we can't arbitrarily scale models without thinking of energy and cost. There is active research on making AI more efficient (model compression, better algorithms), but as of now one could say a limitation is that "Some AI models require substantial computational power and energy resources, posing environmental and financial concerns" [9].

Given these limitations, an overarching theme emerges: Current AI systems are narrow specialists without a deeper understanding or adaptability. They do specific tasks in controlled conditions very well but break down outside those conditions, cannot explain themselves well, and cannot autonomously transfer their learning to radically new tasks in the way humans can. They also lack the emotional and ethical judgment that humans apply in many decisions.

Recognising these limitations is crucial for responsible use of AI. It guides us to keep humans in the loop in critical applications, to use AI for what it is good at (data-driven pattern recognition and automation) and not for what it is bad at (open-ended judgment, understanding context, making ethical decisions). It also directs research: huge efforts are under way in the community to address interpretability, fairness, robustness, and generalisation so that future AI might overcome some of these issues.

In the next section, we will synthesise why, given all these weaknesses, AI is not always the correct or complete solution to a problem and how to decide when traditional methods or human-driven approaches are preferable or needed in complement to AI.

6. Why AI Is Not Always the Right Solution

Artificial Intelligence, for all its remarkable achievements, is not a magic wand suitable for every problem. In concluding this report, we reflect on why AI should be applied judiciously and why sometimes a conventional approach or a human-driven process is better. Overreliance on AI without understanding its limitations can lead to negative outcomes or missed opportunities for simpler solutions. In the following, we summarise key points and guiding principles.

6.1. The Risk of Overreliance and Automation Bias

As AI systems become more prevalent and occasionally outperform humans, there is a temptation to entirely eliminate decision making to them. However, as discussed, AI can fail in unanticipated ways – and if humans have become too reliant, these failures can be catastrophic. A cautionary example can be drawn from aviation: sophisticated autopilot AIs fly planes most of the time, which has improved efficiency and safety, yet when something goes wrong, pilots must quickly step in. If pilots become too dependent and let their skills atrophy, they may not respond effectively in an emergency. Similarly, in medicine, a doctor who does not critically follow an AI diagnosis could mismanage a case if the AI is wrong.

The phenomenon of humans trusting AI recommendations even when wrong - because AI is usually right – has been documented [9]. To mitigate this, organisations must ensure that AI is used as a tool *with* human oversight. Explainability and user training can help users know when to ask AI [6]. For instance, if a loan approval AI flags an applicant as high-risk, a loan officer should review the factors (perhaps the AI provides a profile) and use judgment, rather than blindly accepting it. Maintaining healthy scepticism and verifying AI output against common sense or additional evidence is critical.

In scenarios involving life, liberty, or significant rights (e.g., criminal justice, medical diagnosis), fully autonomous AI decisions are ethically problematic at current capability levels. There should always be a human responsible for final decisions. Overreliance can also cause the so-called "human-out-of-the-loop" problem – where no one really understands or monitors what the AI is doing. This was a factor in some financial flash crashes where trading algorithms interacted in unforeseen ways while humans were too removed to intervene in time.

6.2. When Traditional Methods Are Preferable

Sometimes, the complexity of an AI solution is not warranted. A simpler rule-based system or statistical model may suffice and often proves to be more transparent, easier to maintain, and less dependent on large datasets. For example, if a basic logistic regression using three features achieves precision 95%, there may be little practical benefit in deploying a black-box neural network that improves performance to 97%, but requires ten times more data and computational resources and offers far less interpretability.

In domains with regulatory constraints, such as banking and insurance, the ability to explain is often a legal or operational requirement.

In these cases, a modestly performing yet interpretable model may be the only viable option, even if a more complex AI system performs marginally better.

In low-data regimes, sophisticated AI models tend to overfit or generalise poorly. In such contexts, domain expertise and first-principles reasoning can outperform machine learning. For example, in engineering scenarios where only a handful of prototypes have ever been constructed, physics-based simulations and expert intuition are usually more reliable than data-driven approaches. Similarly, traditional software with explicitly coded rules may be preferable in deterministic environments. A rule-based fraud detection system, while less adaptive, will only flag events based on predefined logic, avoiding unpredictable behaviour that a machine learning model might exhibit when encountering rare or anomalous patterns.

Furthermore, deploying AI costs an overhead. The cost and time associated with collecting high-quality data, training models, validating performance, and maintaining systems in production can be significant. If a task can be adequately addressed with a simple script or a mathematical formula, then it is more efficient and appropriate to avoid AI. As the saying goes, 'don't use a cannon to shoot a mosquito' - in some scenarios, AI is exactly that cannon.

6.3. Ethical and Societal Considerations

AI is not just a technical instrument; it is also a sociotechnical system with wide-ranging implications. One prominent concern is the displacement of jobs. As AI automation accelerates, certain roles may become obsolete, raising questions about the future of affected workers. Overreliance on AI can lead to workforce deskilling, as observed in contexts such as aviation and healthcare, where professionals may lose proficiency when routine tasks are consistently delegated to machines. Organisations adopting artificial intelligence (AI) should therefore consider parallel strategies for workforce development, including retraining and upskilling. Ideally, AI should augment human work, taking over dull, dangerous, or highly repetitive tasks, while leaving roles that require creativity, empathy, and nuanced judgment to humans.

There is also a risk of AI misuse or over-extension — the deployment of AI in domains where its application may be technically possible but ethically inappropriate. For example, the city-wide implementation of facial recognition technologies can help law enforcement, but without appropriate safeguards, such systems risk infringing on civil liberties and perpetuating biased enforcement practices [3]. In such cases, the responsible course may be to forgo AI deployment altogether, recognising that not all problems require algorithmic solutions.

Another key limitation is the lack of nuance in AI-driven decision making. Life-affecting decisions such as parole eligibility, job interviews, or university admissions involve context-sensitive evaluations and moral considerations that cannot be fully encapsulated by the objective function of an algorithm. AI may optimise a narrow definition of success but ignore qualitative and contextual factors to which humans are better equipped to weigh. Human judgment - despite its imperfections - can accommodate values, principles, and empathy in ways that AI systems cannot.

As an industry leader cautioned: 'you need to be careful not to stop at every shiny object out there... and you cannot just drop everything [else]' [3]. In other words, the existence of a sophisticated AI tool does not necessarily mean that it is the right tool for all problems. Maintaining a critical perspective and ethical awareness is essential to ensure that AI serves human values rather than displacing them.

6.4. Building Resilient, Hybrid Approaches

Given the well-documented limitations of AI, a prudent strategy in many applications is to adopt a hybrid approach — one that combines AI systems with rule-based components and human oversight. For example, a medical diagnosis platform might employ AI to analyse imaging data, a symbolic knowledge system to cross-reference symptoms with known conditions, and a physician to make the final decision. In financial services, anomaly detection algorithms can flag suspicious transactions, which are then reviewed by human investigators.

These hybrid designs aim to capture the strengths of each component: the scalability and efficiency of AI, and the contextual awareness and ethical judgment of human operators. In addition, such systems offer resilience. When an AI model encounters a low confidence scenario, it can defer the decision to a human expert, a model design principle commonly referred to as *human-in-the-loop* or a *rejection option*.

Robust AI deployments should also incorporate explicit *failsafes*. For example, if a self-driving vehicle encounters environmental conditions outside its training distribution, such as extreme weather or novel road configurations, it should default to a minimal risk mode or return control to the human driver. Some real-world incidents involving autonomous systems have been attributed to the failure to detect operational uncertainty. Such failure modes can be mitigated by designing systems that explicitly detect and flag unfamiliar or high-risk input.

A related design philosophy is that of *incremental adoption*. Rather than fully automating complex processes in a single step, AI capabilities can be gradually introduced. An initial deployment might operate in a purely advisory role while humans remain in full control. The performance of the system can then be monitored and validated in real-world conditions before expanding the autonomy of the AI. This staged integration builds trust and provides opportunities to refine system behaviour prior to critical reliance.

Hybrid approaches are not merely a stopgap—they represent a principled framework for deploying AI in high-stakes environments. By combining automation with human oversight and procedural safeguards, they offer a pathway to safer, more trustworthy, and more responsible AI systems.

6.5. Staying Aware of AI's Limits

Continued education and awareness are essential for responsible use of AI. Stakeholders, including developers, managers, policy makers, and end-users, must understand both the capabilities and limitations of the technology. It is encouraging that many organisations now establish ethics boards or AI governance frameworks. These typically require testing systems for bias, fairness, and robustness prior to deployment, as well as ongoing monitoring to detect performance drift or emerging failure modes. In some sectors, regulators require model validation and auditability to ensure accountability in algorithmic decision making [9].

AI is not a panacea. It performs exceptionally well on well-defined, data-rich tasks, but falters when faced with ambiguity, context, or value-laden decisions. Overreliance on artificial intelligence introduces new risks. Therefore, responsible deployment requires adherence to several principles:

- Use AI to *enhance* human capability rather than simply replace it. Human–AI teams often outperform either alone.
- We prefer simpler, interpretable methods when they achieve adequate performance. Complexity for its own sake is counterproductive.
- Maintain human oversight, particularly in high-stakes domains, to intervene when AI fails or encounters edge cases.
- Continuously validate, monitor, and train AI systems: these are not 'set-and-forget' tools, but evolving systems.
- Assess broader implications, including ethical, legal, and social dimensions, before adopting AI for a given task.

By remaining aware of the limitations of AI [3], we can make more informed decisions about when to rely on it and when to defer to human judgment or established traditional methods. As an expert aptly observed: *'AI is a tool, not a mentor'* [6]. It is best used in service of human goals guided by human wisdom. With this balanced perspective, we can harness AI capabilities while avoiding overreach, ensuring that technology remains an asset to society, rather than a liability.

7. Contact Novalytics for More Information

Novalytics provides strategic advisory services in information governance, digital transformation, and data strategy for SMEs in regulated and high-risk sectors. We support organisations in modernising their operations through secure, privacy-preserving technologies—ensuring innovation is aligned with regulatory compliance, ethical standards, and long-term resilience.

For expert guidance on digital strategy, transformation planning, or information governance frameworks, please contact us at:

- Website: https://www.novalytics.com
- Email: contact@novalytics.com

References

[1] A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. LIX, no. 236, pp. 433–460, 1950.

[2] J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon, "A proposal for the dartmouth summer research project on artificial intelligence, august 31, 1955," 4, vol. 27, Dec. 2006, p. 12. DOI: 10.1609/aimag.v27i4.1904. [Online]. Available: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1904.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[4] D. Silver, A. Huang, C. Maddison, *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, pp. 484–489, Jan. 2016. DOI: 10.1038/nature16961.

[5] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[7] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[8] T. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[9] J. Jumper, R. Evans, A. Pritzel, *et al.*, "Highly accurate protein structure prediction with alphafold," *nature*, vol. 596, no. 7873, pp. 583–589, 2021.

novalytics