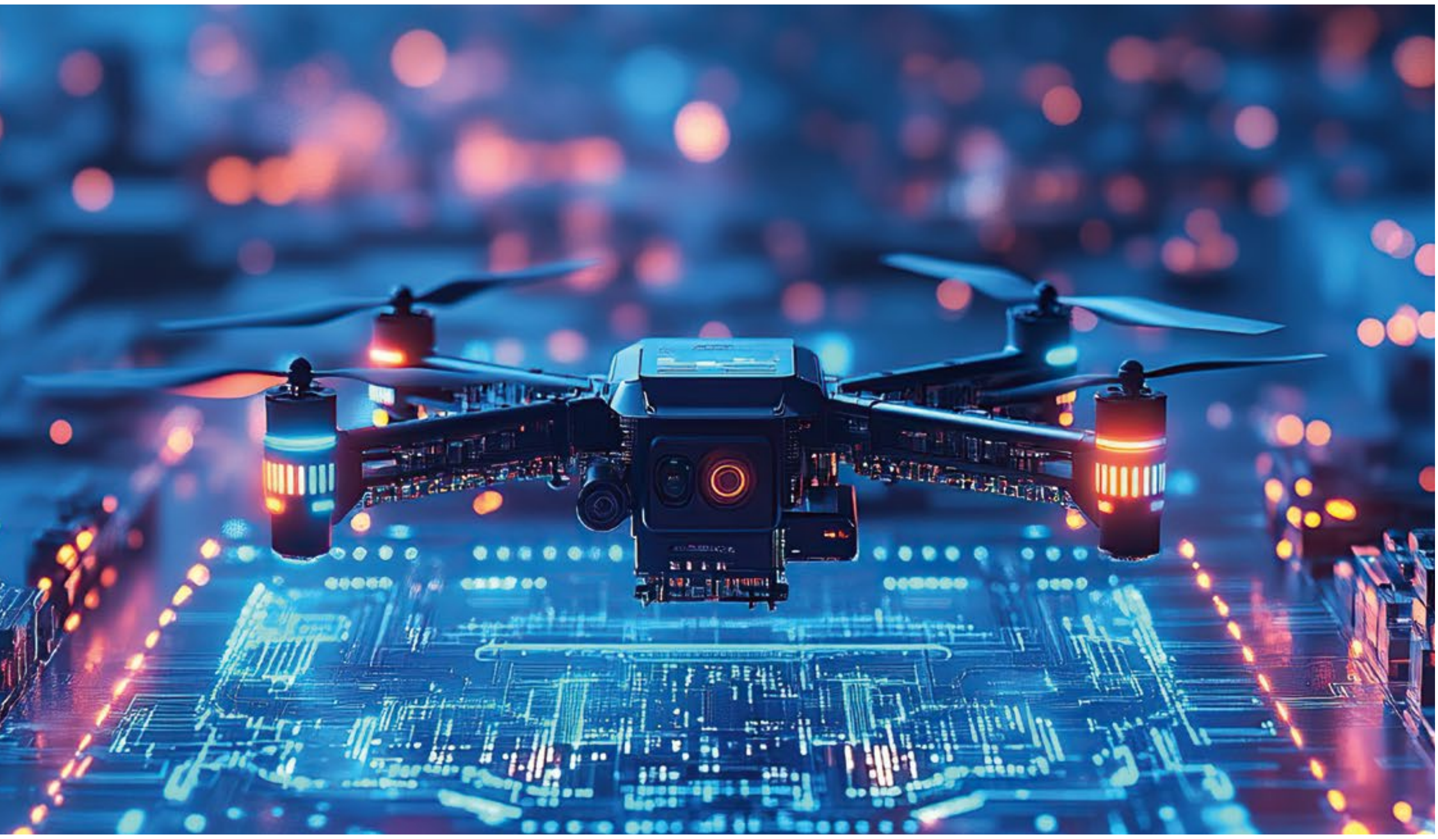
March **2026**



The JAPCC AI Handbook

Practical Considerations for the Warfighter

www.japcc.org

**Joint Air Power
Competence Centre**

Cover Picture: M © Selvi – stock.adobe.com (Generated with AI)

Authors

Colonel (ret.) Antonios Chochtoulas, GRC Air Force
Lieutenant Colonel Nakul Nayyar, CAN Air Force, JAPCC
Major Lucas J. Stensberg, US Space Force, JAPCC
Ms Laura Samsó Pericón, Airbus

Disclaimer

This paper is a product of the JAPCC. It does not represent the opinions or policies of the North Atlantic Treaty Organization (NATO) and is designed to provide an independent overview, analysis and food for thought regarding possible ways ahead on this subject.

Terms of Use – Alteration, Notices

Release

This document is approved for public release, and is distributed to NATO Commands, Nations, Ministries of Defence, and relevant Organisations. Portions of the document may be quoted without permission, provided a standard source credit line, 'Courtesy of JAPCC' is included.

Comments and queries on this document should be directed to C5ISR and Space Branch, JAPCC: c5isr_s@japcc.org

Published and distributed by:

The Joint Air Power Competence Centre
von-Seydlitz-Kaserne
Römerstraße 140
47546 Kalkar
Germany

Telephone: +49 (0) 2824 90 2201
Facsimile: +49 (0) 2824 90 2208
Email: contact@japcc.org
Website: www.japcc.org

M Denotes images digitally manipulated

Follow us on Social Media

From:

The Assistant Director of the Joint Air Power Competence Centre (JAPCC)

Subject:

The JAPCC AI Handbook – Practical Considerations for the Warfighter

Distribution:

All NATO Commands, Nations, Ministries of Defence, and Relevant Organisations

Using this Handbook: The AI Handbook is intended as a practical resource for NATO military leaders and commanders, enabling them to develop a clear understanding of what Artificial Intelligence (AI) is, how it can be applied in military contexts, and the challenges it presents.

While the handbook does not aim to turn its readers into technical experts, it equips them with the essential knowledge required to make informed decisions about adopting and deploying AI responsibly.

By explaining the fundamentals of machine learning (ML) and model-building principles, the handbook helps commanders understand how AI systems learn, process data, and generate outputs.

This understanding is critical when evaluating AI-enabled tools or systems offered by vendors or internal teams, ensuring that decisions are grounded in realistic expectations rather than marketing claims.

The handbook also provides insights into current and emerging military applications of AI, including intelligence, surveillance, and reconnaissance (ISR), autonomous systems, logistics optimisation, and cyber defence. Commanders can use this knowledge to identify opportunities where AI can enhance operational efficiency, improve the speed of decision-making, and reduce human workload. For example, understanding AI-enabled predictive maintenance can help optimise equipment readiness and resource allocation, while awareness of AI-driven ISR capabilities can inform strategic planning.

Equally important, the handbook addresses the limitations and risks of AI, including data bias, algorithmic opacity, and automation bias. These challenges can have serious consequences in high-stakes environments, such as misidentifying targets or the unintentional escalation of conflict. By recognising these risks, commanders can implement safeguards such as human-in-the-loop oversight and rigorous validation protocols. This ensures that AI remains a tool that supports human judgment rather than replacing it entirely.

Ethical and legal considerations are another critical aspect addressed in the handbook. Military leaders must ensure compliance with international humanitarian law and NATO's ethical principles. The guidance on maintaining meaningful human control, accountability, and transparency helps commanders shape procurement decisions, operational planning, and rules of engagement in ways that align with NATO values and global norms.

JOINT AIR POWER COMPETENCE CENTRE

Joint Air Power Competence Centre | von-Seydlitz-Kaserne | Römerstraße 140 | 47546 Kalkar | Germany
Phone: +49 (0) 2824 90 2201 | Email: contact@japcc.org | www.japcc.org

Beyond operational use, the handbook encourages leaders to foster a culture of responsible AI adoption. This includes promoting training programs for personnel, establishing clear protocols for AI deployment, and engaging with international initiatives on AI governance. By doing so, commanders can build trust in AI systems and ensure that their integration strengthens mission effectiveness without compromising safety or ethics.

Finally, the handbook serves as a strategic resource for future readiness. Understanding AI trends and limitations allows commanders to anticipate adversarial use of AI, contribute to NATO's collective security posture, and advocate for international standards that reduce risks associated with uncontrolled proliferation. In short, this handbook empowers military leaders to understand more practically how to employ AI and ML, ask the right questions, demand transparency, and make decisions that balance innovation with responsibility, ensuring that AI becomes an asset rather than a liability in modern warfare.

Vito Cracas
Colonel, ITA AF
Assistant Director, JAPCC

This Page Intentionally Left Blank.

Table of Contents

Executive Summary

Chapters

Annexes

Executive Summary

Artificial Intelligence (AI) is no longer a speculative or emerging technology; it is already shaping the character of modern warfare. The speed, scale, and complexity of contemporary military operations, particularly within Multi-Domain Operations (MDO), exceed human cognitive limits when addressed through traditional manual processes alone. AI-enabled systems offer the potential to process vast volumes of data, identify patterns at machine speed, and support faster, more informed decision-making. At the same time, AI introduces new technical, operational, ethical, and legal challenges that military leaders must understand in order to employ these capabilities responsibly and effectively.

The JAPCC AI Handbook: Practical Considerations for the Warfighter is designed to bridge the gap between highly technical AI literature and the practical needs of military commanders, staff officers, and decision-makers. It does not seek to turn its audience into AI engineers. Instead, it equips leaders with sufficient conceptual understanding to ask the right questions, set realistic expectations, evaluate AI-enabled systems, and integrate AI into military operations without undermining accountability, legality, or trust.

Purpose and Scope

The handbook addresses a critical need within NATO and partner nations: enabling informed leadership decisions on AI adoption in an environment characterised by rapid technological change, increasing data saturation, and accelerating decision cycles. Many AI initiatives fail not because of technical shortcomings, but because decision-makers lack a clear understanding of what AI can and cannot do. This handbook therefore focuses on:

- Explaining core AI and machine learning (ML) concepts in accessible, non-mathematical terms.
- Demonstrating how AI systems are developed, trained, evaluated, and deployed.
- Highlighting realistic military use-cases across intelligence, operations, logistics, cyber, and autonomous systems.
- Identifying limitations, risks, and failure modes inherent to AI-enabled decision support.
- Addressing ethical, legal, and governance considerations central to NATO values and international law.

By grounding AI discussion in operational realities rather than hype, the handbook enables leaders to distinguish between credible capability and marketing-driven claims.

Understanding AI as a Military Tool

A central theme of the handbook is that AI is best understood as a data-driven decision-support tool, not an autonomous replacement for human judgment. Contemporary military AI systems are overwhelmingly examples of Narrow AI: systems optimised for specific tasks such as image recognition, anomaly detection, language translation, or predictive maintenance. While concepts such as Artificial General Intelligence (AGI) and superintelligent AI attract public attention, they remain theoretical and are not relevant to current operational planning.

The handbook explains how modern AI systems, particularly those based on ML and deep learning, derive their capabilities from data rather than explicit programming. This distinction has profound implications for military use. AI performance depends directly on data quality, representativeness, and relevance to the operational environment. As a result, AI systems can reflect biases, amplify errors, or fail unpredictably when exposed to conditions outside their training data.

Understanding this dependency allows commanders to better assess risk, demand transparency from vendors, and avoid over-reliance on automated outputs.

From Concept to Capability: The Machine Learning Pipeline

To demystify AI development, the handbook introduces the ML pipeline, a structured end-to-end process that transforms raw data into an operational capability. This includes:

- Defining the operational problem and determining whether AI is an appropriate solution.
- Collecting, labelling, and preparing data suitable for modelling.
- Selecting and training models aligned with mission requirements.
- Evaluating performance using meaningful operational metrics.
- Deploying models responsibly and monitoring them over time.

This framework highlights that AI success is as much an organisational and human challenge as a technical one. Domain expertise, interdisciplinary collaboration, and sustained oversight are essential. Military personnel, often serving as domain specialists, play a decisive role in shaping AI systems that are operationally relevant, trustworthy, and aligned with commander intent.

Operational Opportunities and Risks

The handbook surveys current and emerging military applications of AI, including data fusion, ISR, autonomous systems, logistics optimisation, cyber defence, and decision-support for command and control. Within these processes, AI can enhance speed, reduce workload, and enable decision advantage when employed appropriately.

However, the handbook gives equal emphasis to limitations and risks, including:

- Data bias and incomplete situational representation.
- Algorithmic opacity and limited explainability.
- Automation bias and over-trust in machine outputs.
- Vulnerability to adversarial manipulation and deception.
- Challenges in testing, validation, and certification (especially for adaptive systems).

These risks are not theoretical. In military contexts, they can contribute to misidentification, escalation, or unintended operational consequences. The handbook therefore stresses the importance of human-in-the-loop or human-on-the-loop control, rigorous validation, and conservative assumptions when deploying AI in high-stakes environments.

Ethics, Law, and Responsible Use

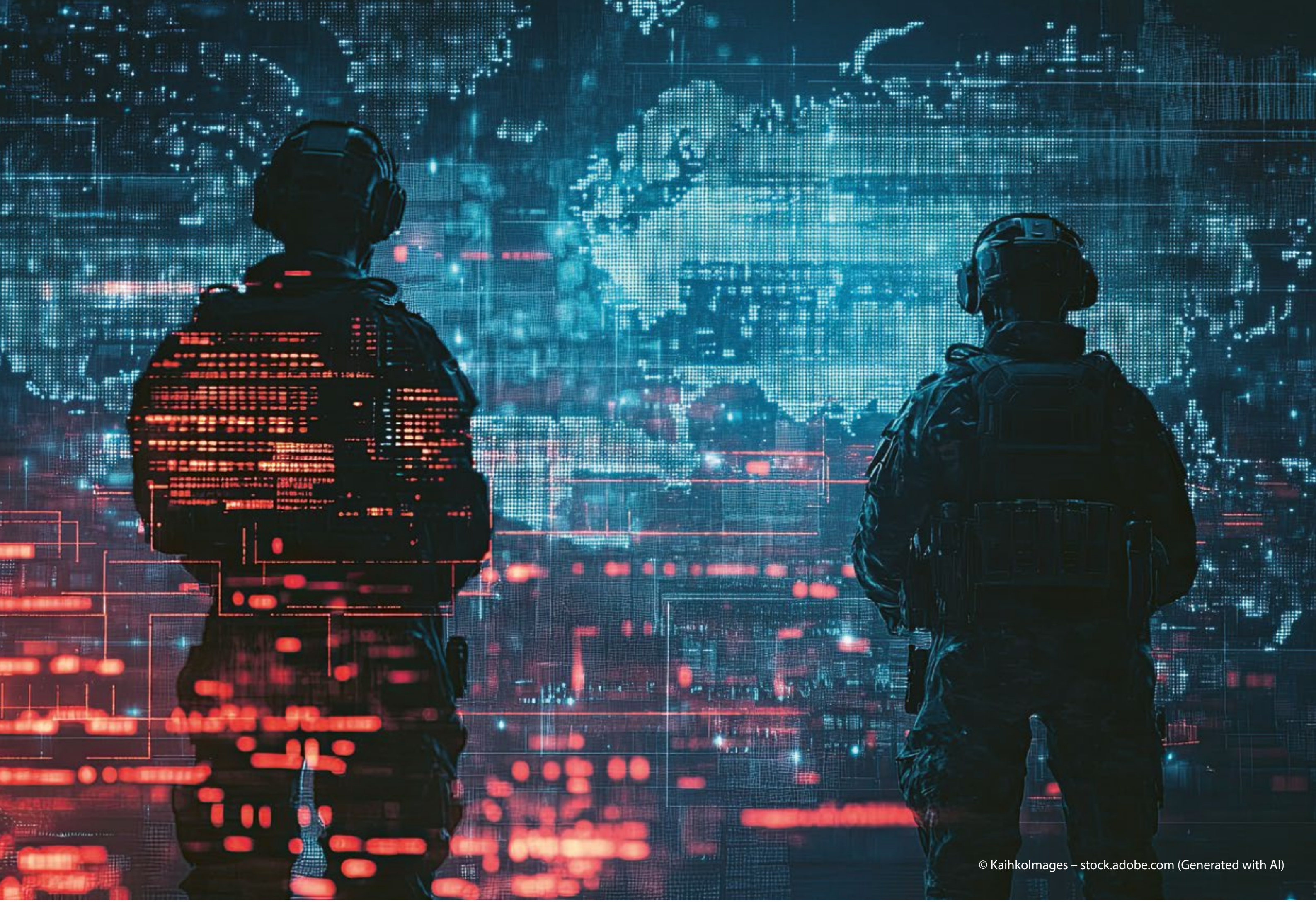
AI adoption in military operations cannot be separated from ethical and legal obligations. The handbook reinforces that compliance with International Humanitarian Law (IHL), NATO principles, and national legal frameworks remains the responsibility of human commanders. AI systems do not bear accountability; humans do.

Key ethical considerations addressed include meaningful human control, transparency, accountability, proportionality, and the dual-use nature of AI technologies. The handbook situates military AI within ongoing international discussions on governance and norms, underscoring NATO's role in promoting responsible use while maintaining strategic advantage.

Strategic Value

Beyond immediate operational utility, the handbook positions AI literacy as a strategic imperative. Adversaries are actively developing and exploiting AI-enabled capabilities, including disinformation, autonomous systems, and cyber operations. A failure to understand AI, both its power and its limits, risks strategic surprise and loss of credibility.

By fostering informed leadership, organisational learning, and realistic expectations, this handbook supports NATO's long-term readiness. It empowers commanders to integrate AI as a force multiplier rather than a liability, ensuring that innovation proceeds in step with responsibility, legality, and operational effectiveness.


© KaihkoImages – stock.adobe.com (Generated with AI)

Chapter 1

The Evolution of AI

1.1 Introduction

Since the dawn of the Information Age, scientists have envisioned what a thinking machine would be like. Pioneers in the 1960s and 1970s employed computers and experimented with mathematical and statistical methods to imitate human intelligence. Their aim was to create tools capable of solving problems faster and more efficiently than humans, eventually leading to what we now call Artificial Intelligence (AI). Their efforts were sometimes successful, generating enormous expectations among stakeholders, although there were periods marked by pessimism and stagnation, such as the 'AI Winter' in the 1980s.[1] Finally, the exponential increase in computing power in the 1990s and 2000s, coupled with the rise of the Internet, enabled researchers to achieve remarkable breakthroughs, including Neural Networks, Machine Learning (ML), and, more recently, the emergence of Generative AI. Many suggest that we are witnessing another technological revolution, the AI Revolution, which is poised to have a transformative and disruptive impact on our societies and fundamentally change the way we interact and work in the years to come.

Could AI change the way we conduct warfare? It is now widely recognised that maintaining superiority in modern warfare necessitates the integration of AI to address numerous challenges. First, the emerging concept of Multi-Domain Operations (MDO) may require AI to synchronise efforts across land, sea, air, space, and the cyber domains, ensuring strategic and tactical advantages. Additionally, the vast number of sensors on the battlefield and the shift towards network-based operations demand sophisticated AI systems capable of processing and analysing enormous amounts of data in real time, a task that now exceeds human cognitive capacity. Furthermore, AI will enable the development of decisive rapid-response capabilities, which are crucial to counter threats at the speed of relevance, while AI-driven strategies are essential to manage threats such as drone swarms effectively. Finally, autonomous systems are evolving rapidly and appear likely to dominate the future battlefield. AI will drive these Human-On-The-Loop (HOTL) or even Human-Out-Of-The-Loop (HOOTL) systems, allowing them to operate in contested environments far behind enemy lines with minimal or no communication with the home base.

1.2 Aim

Non-experts often feel overwhelmed by the myriad terms and technologies associated with AI projects. This is particularly true for military personnel within NATO and national entities, who are tasked with studying and integrating AI into their operations yet frequently lack the requisite expertise. In this context, it is essential for military professionals to develop a clear understanding of AI concepts, tools, and best practices to ensure mission effectiveness and operational success.

Although numerous AI guides and technical papers are available, they tend to be either too general, industry-focused, or excessively technical, making them difficult for non-specialists to comprehend and employ. This handbook seeks to bridge that gap by offering AI insights specifically tailored to the needs of military personnel. It concentrates on key AI concepts and considerations directly relevant to their roles, thereby enabling them to produce higher-quality outcomes that enhance mission success.

1.3 Definition of AI

Before defining AI, it is important first to consider the meaning of intelligence itself. Intelligence is a complex, multifaceted concept without a universally accepted definition. It encompasses various forms, including emotional, verbal, spatial, logical, artistic, and social dimensions, and can be understood within binary, continuum, or multidimensional frameworks. Intelligence can be regarded as a 'suitcase word', carrying many diverse meanings that make precise definition challenging, particularly in discussions relating to AI. This complexity leads to differing interpretations depending on the context, as observed when comparing human intelligence with that of other entities. One useful intelligence definition is, 'the ability to perceive or infer information, and to retain it as knowledge to be applied towards adaptive behaviours within an environment or context'.[2]

AI refers to the capacity of computer systems to perform tasks typically requiring human intelligence. Such tasks include problem-solving, learning, language understanding, decision-making, and even visual perception. AI systems achieve this through techniques such as ML, whereby they learn from data, and more advanced ML like Deep Learning, which employs artificial Neural Networks to simulate human-like reasoning.

The field of AI has traditionally focused on two primary objectives: understanding biological intelligence by embedding it in computers and creating programs capable of performing tasks as well as, or better than, humans. The absence of a precise definition of intelligence has arguably facilitated the rapid growth and advancement of AI, driven by both practical and scientific imperatives to progress and innovate. By concentrating on functional outcomes, what a system can do rather than what it might be, researchers were free to pursue diverse, task-specific solutions without being constrained by philosophical debates regarding the attainment of true consciousness.

1.4 Categorisation of AI

AI can be categorised into three levels of capability: Narrow AI, General AI, and Superintelligent AI.[3]

- **Narrow AI,** or Weak AI, performs specific tasks, such as image recognition or language translation. It is already embedded in our daily life, smartphones, vehicles, and computer systems, and on almost any device that interacts with humans nowadays.

- **General AI** also called Artificial General Intelligence (AGI) or Strong AI, still theoretical, would match human intelligence across all domains. Often described as the 'Holy Grail' of AI research and popularised in films, it could one day provide adaptive systems capable of complex planning, synchronisation, and execution with reduced human intervention.

- **Superintelligent AI,** also theoretical, would exceed human intelligence in all areas, including reasoning, creativity, and strategic decision-making. While speculative, it has already raised significant ethical and security concerns, as such systems could alter the balance of military power and create risks that are extremely difficult to control.

Each category reflects a different level of complexity and a progressively greater potential impact on defence operations, governance, and wider society.

1.5 Different Approaches to AI: Symbolic vs Subsymbolic AI

Since the 1950s, researchers have debated the best approach to developing AI. Some mathematicians favoured logic and deductive reasoning, while others adopted inductive methods based on statistics and probabilities. Researchers from biology and psychology, meanwhile, sought to replicate brain functions and human cognition to create brain-like programs. These perspectives, later termed Symbolic AI and Subsymbolic AI, produced distinct principles and techniques. Despite continued debate, it remains important to draw on both approaches to advance the understanding and development of general AI.

1.5.1 Symbolic AI, or Rule-based AI, is a foundational approach centred on the manipulation of symbols, words or phrases meaningful to humans, through predefined rules.[4] In this paradigm, intelligence is achieved not by mimicking human brain processes but by creating systems that combine and handle these symbols according to established logic. Early advocates believed that by designing sophisticated symbol-processing programs, it might be possible to achieve general intelligence without directly replicating human cognition.

Emerging from mathematical logic and reflecting conscious reasoning, Symbolic AI relies on explicit, rule-based systems that are transparent and easily understood by humans. Using structured languages and algorithms, it produced early expert systems applied in medicine, law, and other fields, where human expertise could be encoded into rules.

Although later overshadowed by ML and Neural Networks, Symbolic AI remains relevant in research and applications. Its strength lies in the clarity and traceability of its rules, contrasting with data-driven methods that rely on statistical learning. This illustrates a fundamental debate in AI: whether intelligence is best captured through structured symbolic reasoning or through complex, brain-like processes.

1.5.2 Subsymbolic AI, inspired by neuroscience, aims to replicate the intuitive and unconscious aspects of human cognition, such as rapid perception for recognising faces or understanding speech. These systems rely on complex mathematical models, particularly Neural Networks, which learn from data without requiring explicit, human-understandable instructions.[5]

An early example was the Frank Rosenblatt *perceptron* (late 1950s), the first Neural Network unit, which laid the foundations for modern Deep Learning.[6] Unlike Symbolic AI's structured rules, the perceptron was designed to simulate brain functions such as pattern recognition.

This shift to Subsymbolic methods marked a significant advance, allowing systems to perform complex tasks through layered numerical computation rather than explicit logic. It reflected a broader trend in AI towards models that learn and adapt from data, embodying more of the implicit processes seen in human intelligence. The success of Neural Networks in image and speech recognition highlights the potential Subsymbolic approaches to tackle problems too intricate for symbolic methods alone.

1.6 The AI Winters

In the 1970s, early advances such as expert systems and rule-based algorithms generated optimism that AI was on the verge of major breakthroughs. This optimism was fuelled in part by researchers who, eager to secure funding and support, made exaggerated claims about the capabilities of AI systems. Such promises fostered the belief that AI was close to achieving human-level intelligence or solving complex operational problems.

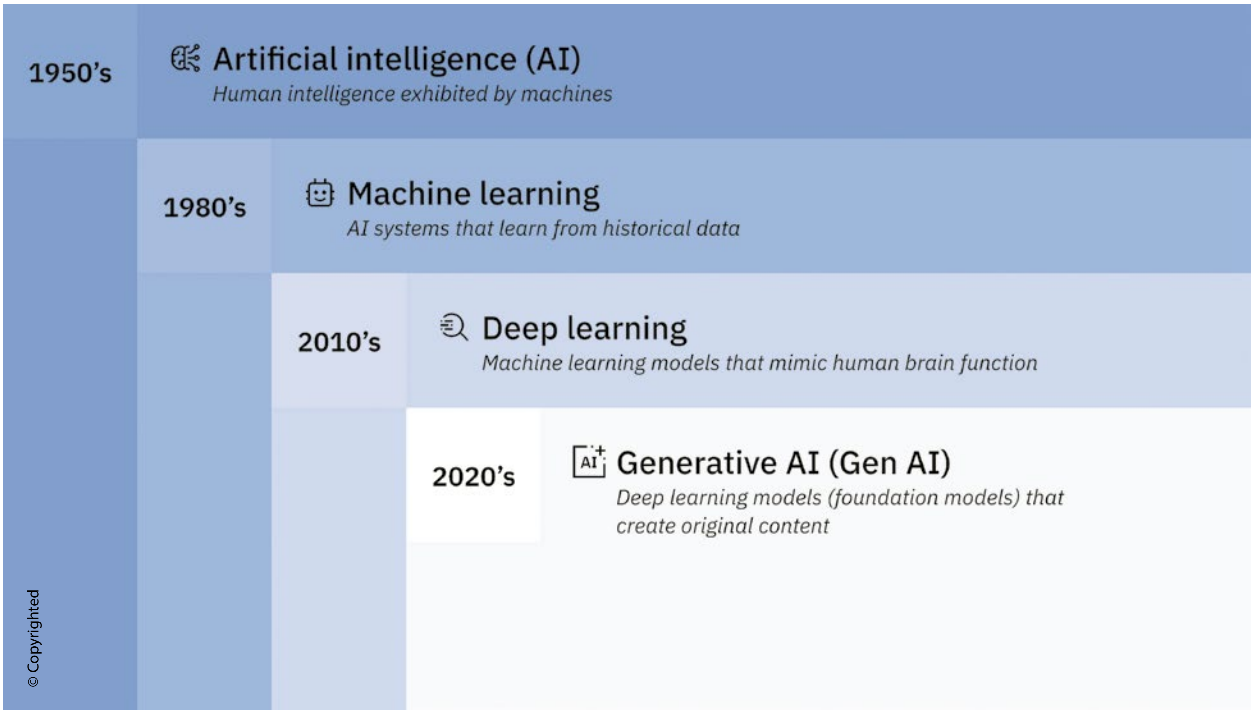
When these systems failed to deliver, and critical assessments exposed their limitations, disappointment grew. Funding declined, and available hardware could not keep pace with theoretical ambitions. This marked the First AI Winter (mid-1970s to early 1980s), which also affected defence-related research, particularly in autonomous navigation and command-and-control applications.[7]

A similar pattern unfolded in the late 1980s. Unmet expectations and growing scepticism triggered the Second AI Winter, stretching into the mid-1990s. Governments, corporations, and other research sponsors sharply reduced investment, and progress stagnated. The limits of computing power, the difficulty of natural language processing, and the inherent complexity of AI made clear that earlier claims had been overly ambitious. It was only with improvements in Neural Networks and the growth of computing power in the mid-1990s that interest and innovation in AI began to recover.[8]

1.7 ML and Deep Learning

The emergence of ML and Deep Learning after the 1990s built upon earlier work on Neural Networks dating back to the 1960s and 1970s. Research into the perceptron and backpropagation algorithms laid the groundwork for later progress, though these methods initially struggled due to limited computational power and the lack of datasets.

As computing capabilities expanded and greater volumes of data became available in the 1990s, ML centred on data-driven pattern recognition, began to flourish, and interest in Neural Networks resurfaced.[9]


1950's
Artificial intelligence (AI)
Human intelligence exhibited by machines
1980's
Machine learning
AI systems that learn from historical data
2010's
Deep learning
Machine learning models that mimic human brain function
2020's
Generative AI (Gen AI)
Deep learning models (foundation models) that create original content
© Copyrighted

A broad overview of AI's evolution from the 1950s to today.

In the 2000s, this renewed momentum gave rise to Deep Learning, a subset of ML employing multi-layered Neural Networks. Deep Learning techniques such as Convolutional Neural Networks (CNNs), combined with hardware advancements like Graphic Processor Units (GPUs), enabled major AI-driven transformations. Technology companies personalised services, healthcare benefited from medical breakthroughs, and the finance sector improved fraud detection. Together, ML and Deep Learning reshaped entire industries, establishing AI as a critical driver of technological development.[10]

1.8 Latest Developments in AI

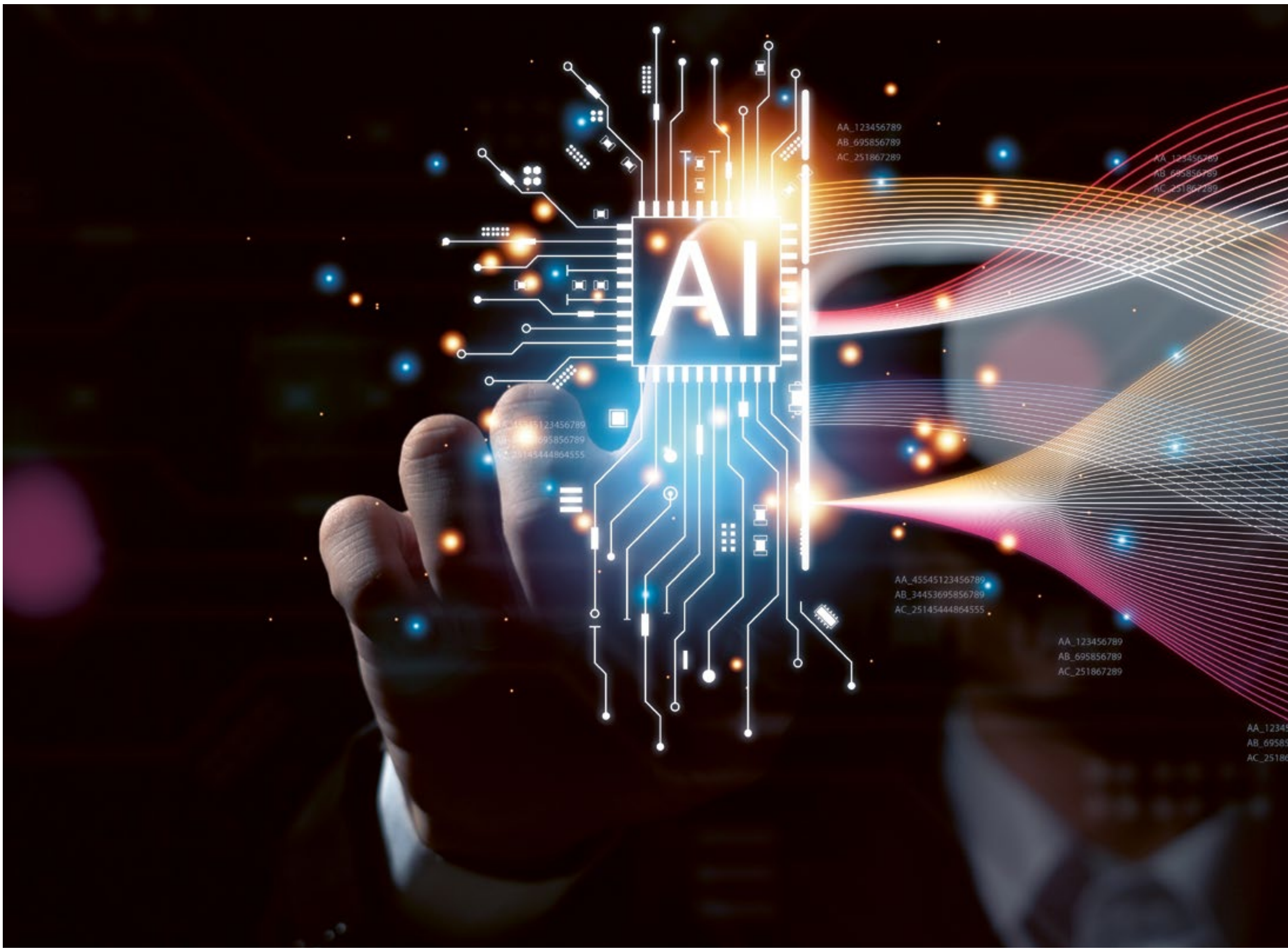
Since 2022, the rapid rise of Generative AI has introduced powerful tools that have drawn global attention for their capabilities in natural language processing and human-like responsiveness. Put simply, Generative AI produces new content such as text, speech, images, or video. Beyond these applications, image and video generation models are transforming creative industries by automating tasks in art, music, and literature. In healthcare, Generative AI contributes to drug discovery and medical imaging, supporting improved diagnostics and treatment options. These advances hold promise for better medical outcomes and overall quality of life. At an individual level, Generative AI is reshaping how people work, communicate, and interact, reinforcing its role as a disruptive technology.

However, its rise also presents significant challenges in ethics and governance. The risks of misinformation, fraud, and privacy breaches through realistic deepfakes and synthetic media are substantial. While these technologies enhance productivity and enable new forms of expression, the automation of roles traditionally carried out by humans raises concerns over job displacement. This highlights the urgent need for reskilling and the creation of new professional pathways. Furthermore, the speed of AI development is outpacing regulatory frameworks, underlining the importance of strong governance to manage its legal and ethical implications.

1.9 Conclusion

The history of AI is one of ambition, challenge, and exponential growth. From Symbolic and Subsymbolic approaches to the cycles of the AI Winters, this evolution provides essential perspectives for separating hype from reality. Such foundational knowledge is not merely academic: it is necessary for making informed, strategically sound decisions about integrating AI into our military structures, which will be the central focus of the chapters that follow.

1. IBM. (n.d.). The history of AI. IBM. Retrieved 9 September 2025, https://www.ibm.com/think/topics/history-of-artificial-intelligence.
2. Silvercrest Group. (n.d.). AI—Answering a few key questions. Silvercrest Group. Retrieved 9 September 2025, https://www.silvercrestgroup.com/artificial-intelligence-answering-a-few-key-questions/.
3. Syracuse University iSchool. (1 April 2025). Types of AI: Explore key categories and uses. Syracuse University iSchool. Retrieved 9 September 2025, https://ischool.syracuse.edu/types-of-ai/.
4. DataCamp. (12 May 2023). What is symbolic AI? DataCamp. Retrieved 9 September 2025, https://www.datacamp.com/blog/what-is-symbolic-ai.
5. Schneppat. (n.d.). Symbolic AI vs. subsymbolic AI. Retrieved 9 September 2025, https://schneppat.com/symbolic-ai-vs-subsymbolic-ai.html.
6. De La Cruz, R. (1 November 2023). Frank Rosenblatt's perceptron, birth of the neural network. Medium. Retrieved 9 September 2025, https://medium.com/@robdelacruz/frank-rosenblatts-perceptron-19fcce9d627f.
7. AI for Everyone. (13 March 2025). AI winter: Why enthusiasm around AI sometimes wanes. AI for Everyone. Retrieved 9 September 2025, https://aiforeveryone.blog/en/ai-for-everyone/ai-winter-why-enthusiasm-around-ai-sometimes-wanes.
8. Ibid.
9. Dataversity. (3 December 2021). A brief history of ML. Dataversity. Retrieved 9 September 2025, https://www.dataversity.net/a-brief-history-of-machine-learning/.
10. Ahmed, S. F. (2025). Unveiling the frontiers of DL: Innovations and applications. Applied Intelligence. Advance online publication. https://doi.org/10.1007/s10489-025-06259-x.



Chapter 2

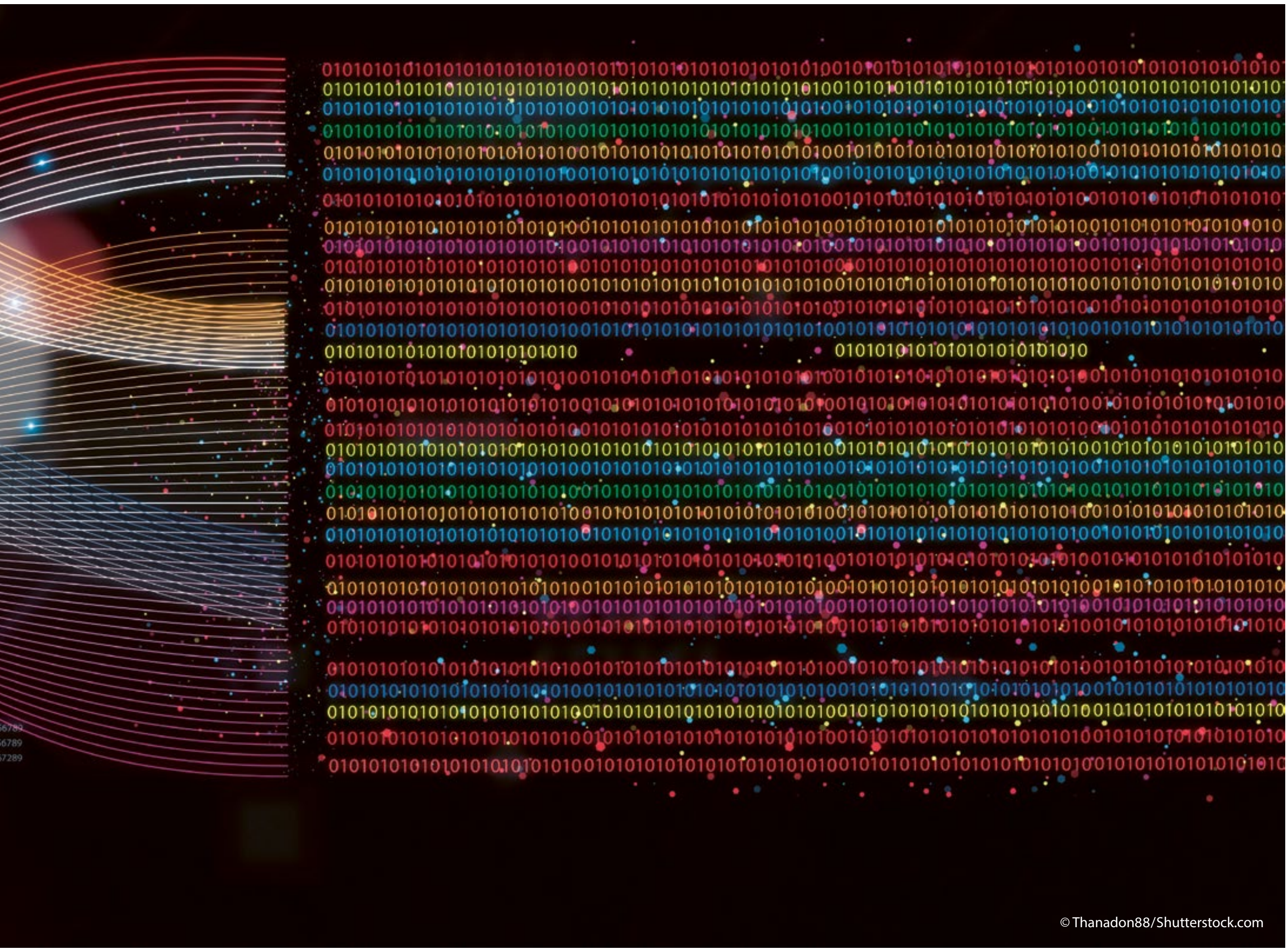
Technology Overview

2.1 Introduction

This chapter examines the technological backbone of modern AI, considering how it has evolved, why certain approaches have become so influential, and what their scope of capability is. Although AI encompasses a wide range of methods, the focus here will remain on Sub-symbolic AI, the branch that has moved to the forefront in recent years. Defined by adaptability and diversity, this approach includes technologies such as Neural Networks and Deep Learning, which have transformed fields from image recognition to natural language processing.

2.2 Machine Learning

Machine Learning (ML) is a subfield of AI that enables computers to learn from data without being explicitly programmed.[1] The roots of ML go back to the 1960s, with IBM's Arthur Samuel and his pioneering work on using ML to play checkers. He demonstrated that a computer could improve its performance beyond the direct knowledge of its programmer.[2] This early breakthrough laid the foundation for a field that now underpins technologies we rely on daily, from fraud detection to vision and speech recognition.



© Thanadon88/Shutterstock.com

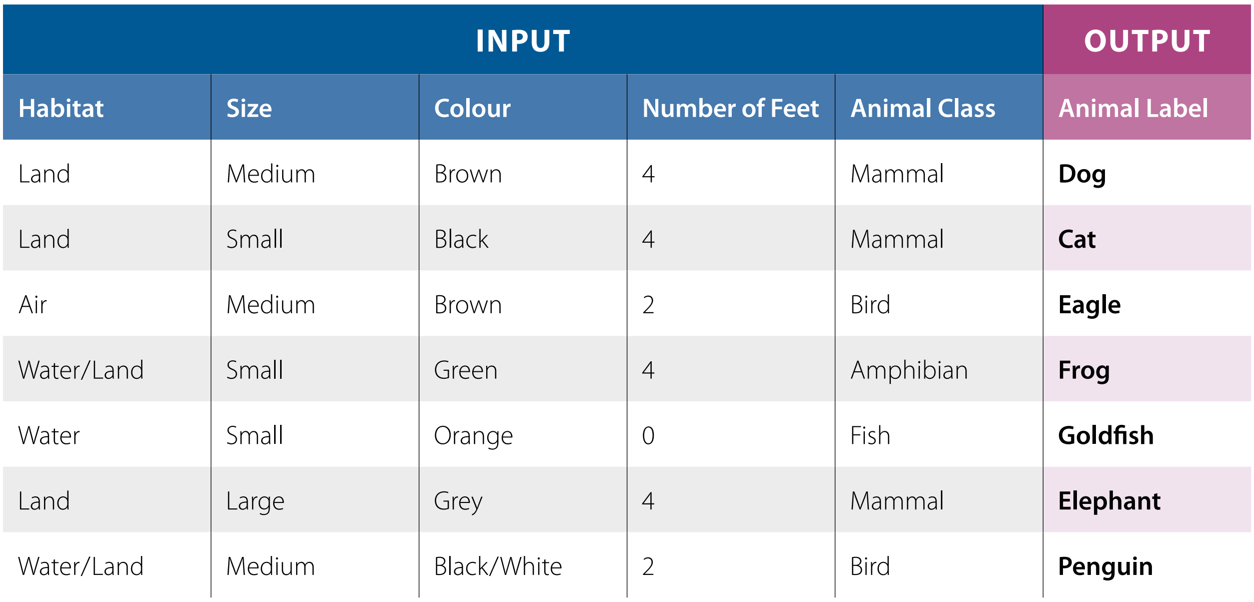
At its core, ML applies statistical models to detect patterns in data and make predictions or decisions. Unlike traditional programming, where explicit instructions are provided for every task, ML algorithms learn from input data, building models that can generalise from examples.

The process usually involves feeding data into an algorithm, training the model, and then testing its performance on new, unseen data. This iterative cycle allows the system to refine predictions and improve over time, mirroring the way humans learn from experience. It is important to note that ML does not rely exclusively on Neural Networks but also employs more classical algorithms such as regression and support vector machines.

2.3 Types of Machine Learning

ML is fundamentally categorised into three main types: supervised learning, unsupervised learning, and reinforcement learning.[3] Each category addresses different kinds of problems and employs distinct methodologies to derive predictions from data.

Supervised learning focuses on learning patterns from labelled data, where each input has a corresponding output. The model is trained by analysing input-output pairs, making predictions, and adjusting based on feedback to reduce errors. During training, data are typically divided into a training set and a test set to evaluate how well the model adapts to new data. This approach requires extensive upfront labelling,



INPUT					OUTPUT
Habitat	Size	Colour	Number of Feet	Animal Class	Animal Label
Land	Medium	Brown	4	Mammal	**Dog**
Land	Small	Black	4	Mammal	**Cat**
Air	Medium	Brown	2	Bird	**Eagle**
Water/Land	Small	Green	4	Amphibian	**Frog**
Water	Small	Orange	0	Fish	**Goldfish**
Land	Large	Grey	4	Mammal	**Elephant**
Water/Land	Medium	Black/White	2	Bird	**Penguin**

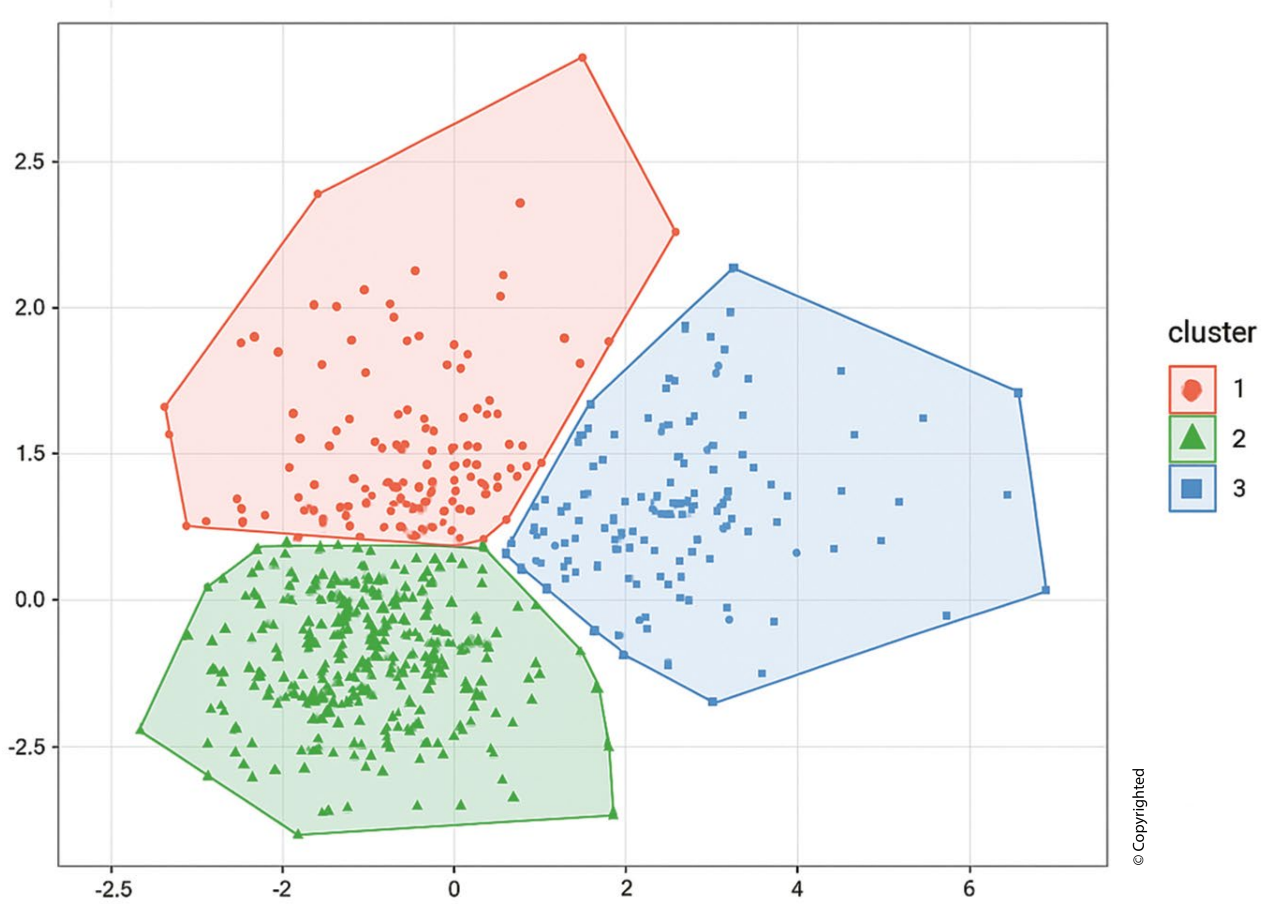
Figure 1: *A sample training table.*

where human intervention provides known data, such as classifying spam emails or, in a military context, identifying enemy tanks versus friendly trucks in satellite imagery based on a pre-labelled dataset. Figure 1 shows a simplified training table that could be used to train a classification algorithm to identify animals based on features like size, colour, and number of feet.

By comparing its predictions to the correct outcomes, we can fine-tune the system to make better predictions over time. Accuracy may also be improved by adding to the training table and supplying a larger number of examples. This approach scales from basic tasks, such as digit recognition, to complex applications, such as fraud detection and image classification. However, supervised learning requires time and expertise to label data accurately, which can make the process labour-intensive and limit scalability.

In contrast, *unsupervised learning* works with unlabelled data to identify hidden patterns or structures without predefined outcomes. The algorithm discovers relationships within the data and groups the examples in a multidimensional space based on similar characteristics or behaviours. The number of dimensions of this space depends on the number of characteristics identified in the data. Unsupervised learning models can be computationally demanding due to the large, unlabelled datasets they handle. Algorithms such as clustering and association analysis help these models find patterns without needing predefined labels. Some human intervention is still required to validate and interpret the results. For example, while the model might identify that online shoppers frequently purchase baby products together, a data analyst would validate that this grouping aligns with practical observations, just as an intelligence analyst might validate that a cluster of maritime vessel activity requires further investigation.

Reinforcement learning represents the most advanced category, where the model learns through continuous interaction with its environment, receiving feedback based on its actions. This iterative process allows the model to optimise its strategy over time. A classic example is training a model to play a video game, where it learns to maximise its score by avoiding negative outcomes and pursuing positive rewards. The same principle is used to train an unmanned aerial system (UAS) to autonomously navigate complex, contested airspace or to optimise electronic warfare responses against an evolving threat.



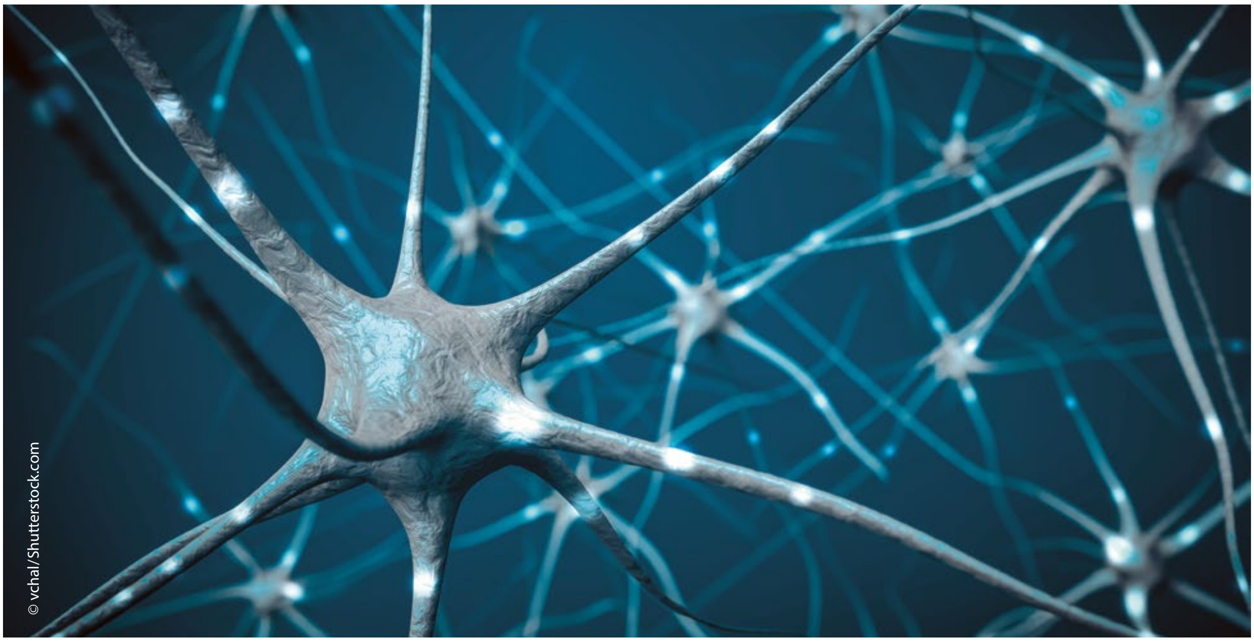
The image above shows a cluster plot created by a 'clustering', an unsupervised learning algorithm. Each observation has been grouped into only one of the three clusters.

Reinforcement learning is widely used in autonomous systems, such as self-driving cars, where the model continuously improves its performance based on real-time data and experience. Typically, the model is trained in a simulated environment resembling the real one, to accelerate the process and minimise potential damage. This is especially relevant for robots or autonomous vehicles, where a negative outcome might have catastrophic consequences.

The choice of ML technique depends on the data structure, volume, and desired outcome. Supervised learning is preferable when labelled data are available and the task has a well-defined output, ensuring high accuracy and reliability. Unsupervised learning is more suitable for exploring data to uncover underlying structures or patterns. Reinforcement learning is best applied to dynamic environments where continual learning and adjustment are required.

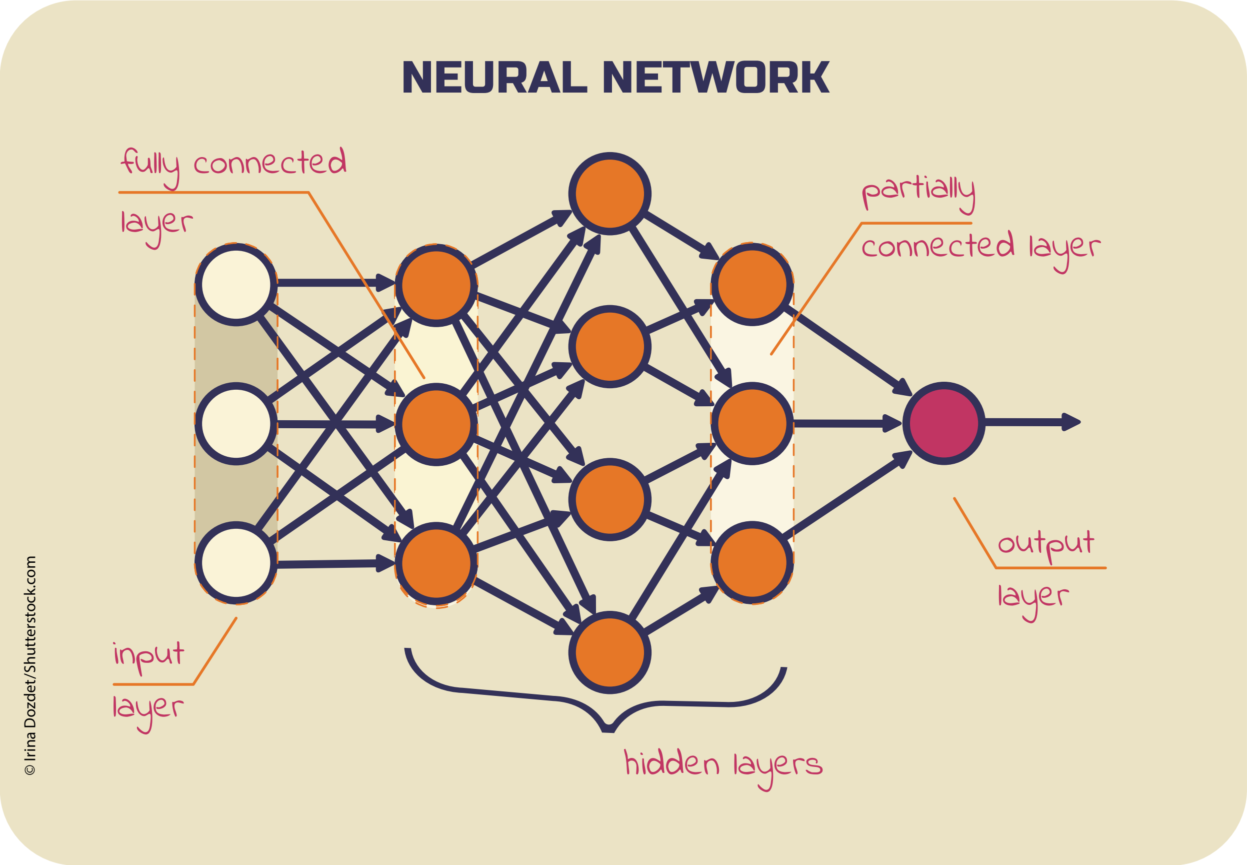
2.4 Neurons and Neural Networks

The human brain contains approximately ten billion neurons, each one typically consisting of a soma, or cell body, dendrites that receive signals, and axons that transmit signals. Neurons communicate through electrochemical pulses. When a pulse crosses a synapse, it can either excite or inhibit the receiving neuron's activity. The cumulative effect of these pulses determines whether the neuron will fire its own signal. It is estimated that the brain contains over half a trillion synapses, enabling complex activities such as perception and reasoning.



© vchal/Shutterstock.com

A sample structuring of the brain's neurons.



© Irina Dozdet/Shutterstock.com

A simple multilayer neural network, showing the flow of data from the input layer, through hidden layers where processing occurs, to the final output layer.

In 1943, Warren McCulloch and Walter Pitts proposed that neurons act as 'logic units' and created models showing that networks of these units could perform any computational operation.[4,5] Their work laid the foundation for viewing neurons as mathematical abstractions with inputs and outputs, capable of complex processing and computation.

Later, in 1959, psychologist Frank Rosenblatt developed the perceptron, an early artificial neuron that could be trained through supervised learning to process data. A decade later, in 1969, Marvin Minsky and Seymour Papert published *Perceptrons,* in which they argued against some of the hubris associated with AI at the time, while indirectly sparking research that would decades later overcome the very limitations they exposed.[6] Namely, it paved the way for the perceptron with an extra layer, a multi-layer neural network, which forms the backbone of much of today's AI technology.[7]

Within these multilayer neural networks are the input layer, hidden layers, and output layer. The input layer receives the raw data the model seeks to interpret and use for prediction. The hidden layers are intermediate stages where neurons apply weights and activation functions to transform and extract meaningful patterns. The output layer produces the final predictive result.[8]

2.5 Deep Learning

Since the 2010s, Deep Learning has emerged as the leading paradigm, often giving rise to the misconception that AI refers exclusively to it. In reality, AI encompasses a broad range of approaches aimed at creating intelligent machines, with Deep Learning being just one method within machine learning, itself a subfield of AI focused on data-driven learning and experience.

Deep Learning evolved as an attempt to make traditional ML techniques more capable. As noted earlier in this chapter, traditional ML involves statistical models and algorithms that require human effort to train the models and evaluate their performance. By contrast, Deep Learning models learn to identify features directly from raw data. They automatically detect complex patterns, particularly in unstructured data such as images, audio, or text, for example, full-motion video from a UAS, intercepted voice communications, or open-source intelligence reports.

The term *deep* refers to the fact that Deep Learning employs Neural Networks typically organised into multiple layers, meaning that computation paths from inputs to outputs involve many steps.[9] Deep Learning is currently the most widely used approach for applications such as visual object recognition, machine translation, speech recognition, speech synthesis, and image generation. However, Deep Learning requires far more data and computational power than traditional ML models.

2.6 Generative AI

Recent advancements in AI have introduced powerful text-generation tools, such as ChatGPT, which fall under the broad category of generative AI. Generative AI refers to systems designed to produce new content, whether text, speech, video, or images, by leveraging learned patterns from vast data sources rather than following step-by-step instructions. At the core of these tools lies Large Language Model (LLM) technology, which applies Deep Learning principles to analyse complex patterns and relationships within extensive text datasets. This technology enables ChatGPT and similar AI systems to predict and generate coherent, contextually appropriate word sequences that mimic the natural flow and nuance of human language, resulting in human-like output.

The operation of LLMs consists of two primary phases: training and generation. Training itself involves two stages. First, the model learns statistical patterns from extensive text datasets and adjusts its billions of internal parameters to develop a general word-prediction capability. Secondly, a fine-tuning process, which incorporates human feedback to model outputs, optimises word-prediction accuracy within given contexts, thereby shaping the models' final form. Once trained, the system applies its acquired knowledge to generate new outputs in response to prompts, refining its output continuously based on preceding content and contextual cues until the desired result or completion condition is reached.

2.7 Multimodal Systems

The latest development in AI are multimodal systems. These are advanced generative AI models designed to process and integrate multiple types of data, such as text, images, audio, and video, to perform tasks that require understanding across different modalities. Unlike unimodal AI systems, which focus on a single data type (for example, language models working only with text), multimodal systems combine information from diverse sources, enabling them to better approximate human-like perception and reasoning.

For example, a multimodal AI system could take a picture to identify objects while simultaneously generating a descriptive caption in natural language, a capability with obvious applications in automated intelligence reporting. It could also interpret video by recognising visual elements, transcribing spoken dialogue, and drawing conclusions from both visual and audio inputs. This ability to process and integrate different forms of data gives multimodal systems a more holistic understanding of their environment, opening up new possibilities for AI applications in areas such as healthcare, autonomous vehicles, entertainment, and human-computer interaction.

In autonomous vehicles, multimodal systems can integrate real-time data from cameras, radar, and GPS to navigate complex environments. They could also revolutionise the way humans interact with computers: typing would be only one of the possible input methods, alongside live video and sound. In entertainment, AI models like OpenAI's GPT-5 and DALL·E exemplify how text-based and image-based inputs can work together, allowing users to generate both written content and visuals from a single prompt.

The development of multimodal systems also marks significant progress towards creating AI that is more closely aligned with human intelligence. Humans constantly process a variety of inputs simultaneously, using multiple senses to build a comprehensive understanding of the world. Multimodal AI seeks to replicate this ability, allowing machines to interact with their environment and with users in more dynamic and intuitive ways.

2.8 Frozen vs. Adaptive AI: Predictability and Performance

The behaviour of an AI system depends significantly on its internal parameters and on the volume and quality of data used during training. After deployment, a system may continue to learn from new data, gathered through sensors or collected at various stages of its lifecycle, thereby refining its performance. However, this ongoing adaptability may introduce challenges in accurately predicting the system's behaviour, particularly in real-time applications. This reality highlights a central dilemma for military use: the trade-off between predictability and performance. A static, predictable system provides high assurance and is easier to certify for critical missions, but it cannot adapt to novel threats. Conversely, an adaptive system offers superior performance and resilience in dynamic environments, yet its evolving nature introduces inherent unpredictability, complicating test, evaluation, verification, and validation (TEVV).

This distinction gives rise to two categories of AI systems, depending on whether their learning and behaviour are fixed after development or permitted to evolve during operational deployment:

Frozen Artificial Intelligence: systems whose behaviour is determined by training data and parameters fixed at the end of the development phase, ensuring they never diverge from the release version.

Adaptive Artificial Intelligence: systems whose parameters continue to adapt during execution, drawing on the knowledge acquired and the experience they gain throughout deployment.

2.9 Impact of Quantum Technologies on AI

Quantum computing is an emerging technology that leverages the principles of quantum mechanics, such as superposition and entanglement, to create states that scale exponentially with the number of qubits, thereby significantly outperforming the capabilities of classical computers.[10] This ability to process vast and complex problems has made quantum computing a promising area of research for both academic and industrial communities.

The integration of AI and quantum computing promises transformative advancements, enabling systems to perform tasks that are currently beyond reach. Quantum computing's ability to handle complex calculations and vast datasets at unprecedented speeds could dramatically enhance AI's problem-solving capacity, particularly in fields such as optimisation, cryptography, and large-scale simulations. Some researchers suggest this might even accelerate progress towards Artificial General Intelligence (AGI), characterised by a more generalised understanding akin to human cognition.

Nevertheless, real-world applications of quantum computing remain at an early stage. Significant technical and theoretical challenges must be addressed before practical quantum computers can be realised, and substantial research is still needed to integrate them effectively with existing AI algorithms. Despite these obstacles, the future of quantum-enhanced AI is highly promising. The nation or alliance that first achieves a functional synthesis of quantum computing and AI is likely to gain a decisive, and potentially irreversible, strategic advantage.

2.10 Fields of Application

The following sections discuss some applications of the aforementioned technologies, with a focus on operational impacts within military contexts.

2.10.1 Natural Language Processing

Natural Language Processing (NLP) enables computers to understand, process, and generate human language, making effective human-machine communication possible. Its applications in the military domain are extensive and potentially transformative, reshaping how humans interact with computers. In general, tasks empowered by NLP could include:

- Speech recognition;
- Text-to-speech synthesis;
- Machine translation;
- Information retrieval and extraction;
- Question answering.

For NATO, its applications are not theoretical but directly address core warfighting challenges in three critical areas:

Intelligence Fusion at Speed of Relevance: Modern battlefields are saturated with unstructured, multilingual data from sources like signals intelligence (SIGINT), open-source intelligence (OSINT), and battlefield reports. NLP could enable AI to process this data at a scale and speed impossible for human analysts, automatically identifying key entities, translating communications, and detecting patterns or shifts in sentiment. This transforms torrents of raw, unstructured data into a coherent, fused intelligence picture.

Accelerating Command and Control (C2): The primary means of human-computer interaction, keyboards, mice, and touchscreens, are slow and inefficient in high-tempo combat. NLP can transform C2 by allowing commanders to issue verbal commands to systems, automatically transcribing voice communications into structured reports, and even analysing commander's intent to assist planners. This reduces the cognitive load, accelerates decision-making, and contributes directly to decision advantage.

Enhancing Alliance Interoperability: Overcoming linguistic diversity has long been one of NATO's most enduring operational challenges. NLP-powered real-time translation and transcription can break down these barriers, enabling, for example, a Polish pilot, a French Joint Terminal Attack Controller, and an American air battle manager to communicate seamlessly. More than mere translation, this ensures shared tactical understanding and dramatically improves the cohesion and lethality of multinational forces.

In addition, because programming languages are structured like human languages, NLP could allow computers to write computer code faster and more accurately than humans, even from spoken instructions. This capability would extend programming to those with no technical background, enabling operators to task computers directly in natural language rather than specialised coding languages.[11]

2.10.2 Computer Vision

Although recognising objects and navigating through environments appear effortless for humans, they are immensely complex for machines. Computer vision, the ability of computers to 'see' their environment, requires advanced algorithms and significant computational resources.[12]

Machines use Neural Networks to identify patterns in images and video streams, extracting the information necessary for navigation and object recognition. In the military domain, computer vision has a wide range of applications, from enabling autonomous systems to ISR and precision targeting. The greatest benefit, however, lies in enhancing machines' ability to perceive their surroundings, allowing them to operate autonomously and produce reliable, accurate results.

Other tasks include labelling images, reconstructing scenes from multiple viewpoints, and controlling movement based on visual input. Furthermore, integrating computer vision with high-performance computation could significantly accelerate image processing, reducing the time required for operators to identify, prioritise, and assign targets on the battlefield.

2.10.3 Robotics/Autonomous Systems

Robots and autonomous systems are physical agents that perform tasks by interacting with the physical world. They are also equipped with diverse sensors, ranging from cameras and microphones to gyroscopes, that allow them to perceive their environment. Currently, robots do not exhibit human-level agency in complex tasks and cannot autonomously navigate their surroundings or apply reasoning to achieve goals independently.

To understand and interact with the world, robots use their sensor measurements to build internal representations of their environment, improving situational awareness. ML plays a critical role in this process by enabling robots to process and identify patterns in sensor data while simultaneously developing an internal model that informs decisions about movement or action.[13]

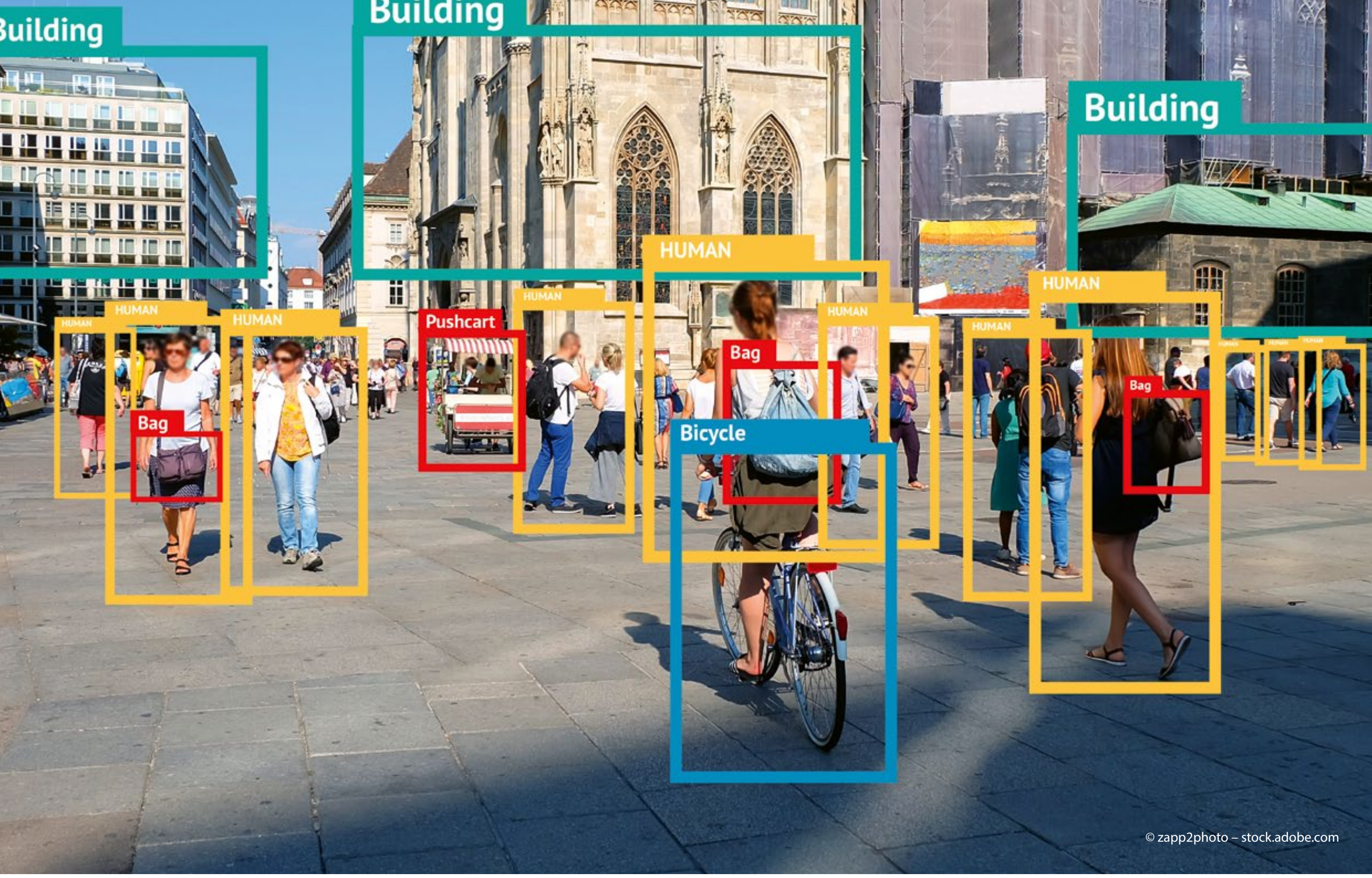
Future autonomous systems will increasingly rely on AI to perform their functions. In military applications, drones across air, land, and sea domains will need to operate in complex and congested environments, potentially without GPS guidance or secure, reliable communication channels. To achieve this, these systems must be AI-enabled. Various subsystems will incorporate AI to support deep learning and NLP capabilities, essential for analysing data and extracting actionable insights. Additionally, computer vision will aid navigation, while AI-driven reasoning will allow autonomous decision-making in real-time.

2.11 Conclusion

Having surveyed the technological landscape of modern AI, a clear picture emerges: these systems are not magical black boxes, but powerful, data-driven tools built on comprehensible principles. For the military professional, understanding the concepts of deep learning, NLP, and the critical assurance implications of frozen versus adaptive systems is essential for responsible adoption.

The next chapter shifts from theory to application, detailing the structured end-to-end process, the machine learning pipeline, required to successfully develop and deploy these capabilities in the real world.

[1] Theobald, O. (2021). *Machine learning for absolute beginners* (3rd ed.). Scatterplot Press, p. 8.
[2] IBM. (n.d.). The games that helped AI evolve. IBM. Retrieved 9 September 2025, https://www.ibm.com/history/early-games.
[3] Ibid 1, p. 15.
[4] Nilsson, N. J. (2009). *The quest for artificial intelligence: A history of ideas and achievements.* Cambridge University Press. https://ai.stanford.edu/~nilsson/QAI/qai.pdf, p. 34.
[5] Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th ed.). Pearson, Chapter 21.
[6] Mitchell, M. (2019). *Artificial intelligence: A guide for thinking humans* (p. 34). Farrar, Straus and Giroux.
[7] Ibid, 1.
[8] IBM. (n.d.). *What are Neural Networks?* Retrieved 9 September 2025, https://www.ibm.com/de-de/think/topics/neural-networks.
[9] Ibid 5, p. 750.
[10] Quantum News. (10 May 2024). *Quantum computing's potential to revolutionize AI: A comprehensive overview.* Quantum Zeitgeist. Retrieved 9 September 2025, https://quantumzeitgeist.com/quantum-computings-potential-to-revolutionize-ai-a-comprehensive-overview/.
[11] Ibid 5, chapter 23.
[12] Ibid 5, chapter 25.
[13] Ibid 5, chapter 26.



© zapp2photo – stock.adobe.com

Chapter 3

Machine Learning Pipeline (Part 1): Data Collection and Preparation

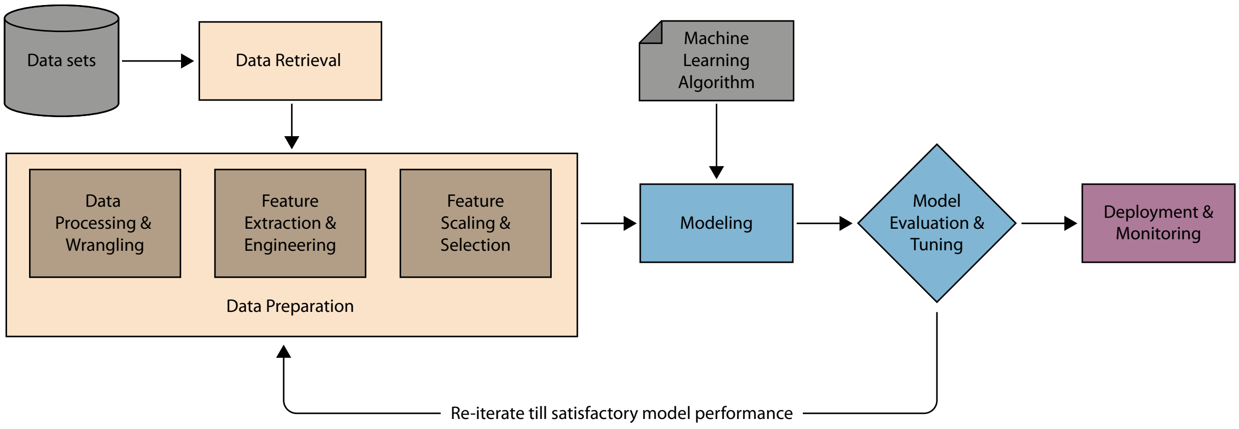
3.1 Chapter Overview

As outlined in Chapter 2, Machine Learning (ML) is a field within Artificial Intelligence (AI) in which computers learn and adapt from data without humans explicitly designing their algorithms. This is achieved through statistical models that detect patterns within data and make predictions based on those patterns. While the number of use-cases for AI/ML is rapidly expanding, there remains a high-level pipeline that can be applied to almost all AI/ML problem sets. This ML pipeline defines the practical steps required to deploy a model that generates new insights from raw datasets.

Chapter 3 focuses on gathering and preparing raw datasets for modelling, while Chapter 4 addresses ML modelling, evaluation, and deployment. Familiarity with technologies introduced in Chapter 2 is important before proceeding to Chapters 3 and 4, as this section transitions from *what* AI/ML is to *how to*, offering actionable steps and practical considerations. For readers without a technical background, some of the details in the next two chapters may be more challenging to follow.

Through its Digital Transformation Vision in late 2022, NATO decided to harness technologies essential to enabling Multi-Domain Operations (MDO). AI and ML fall under this umbrella. Consequently, NATO personnel may increasingly need to engage with disruptive technologies firsthand. Although revolutionary today, such technologies will soon become commonplace. Our adversaries are already leveraging them; we cannot afford to wait.

The aim of the next two chapters is to equip readers with tangible procedures designing their own ML


Data sets
Data Retrieval
Machine Learning Algorithm
Data Processing & Wrangling
Feature Extraction & Engineering
Feature Scaling & Selection
Data Preparation
Modeling
Model Evaluation & Tuning
Deployment & Monitoring
Re-iterate till satisfactory model performance

© Used with permission of Springer Nature BV, from Practical Machine Learning with Python: A Problem-Solver's Guide to Building Real-World Intelligent Systems, Sarkar Dipanjan, Bali Raghav, Sharma Tushar, 1st Edition, 2017; permission conveyed through Copyright Clearance Center, Inc.

A flowchart showcasing a standard machine learning pipeline which will be the foundational structure of the next two chapters.

pipeline within their organisation. This does not imply that an organic NATO or national military organisation will necessarily have all the resources or personnel to build the pipeline independently, but it will at least inform detailed gap analyses. At the very least, domain-specific expertise is essential to constructing an effective ML pipeline. For example, a missile defence Company Commander deeply familiar with his own TTPs might share operational insights with computer scientists specialising in data science and engineering.

Finally, much of the material in this chapter is derived from the ten-week course *Machine Learning and Artificial Intelligence for Cyber Applications,* jointly offered by the NATO School Oberammergau and the US Naval Postgraduate School. The flowchart shown above is from that course, and will drive the structure of the next two chapters.

3.2 Resources

Implementing an ML pipeline is resource-intensive; therefore, not all organisations will be immediately prepared to adopt AI/ML. Successful implementation depends on team expertise and an appropriate information technology (IT) backbone. The following sections outline these elements.

3.2.1 Team Dynamics

There is no one-size-fits-all team for ML, as each problem set can vary significantly in size and complexity. However, certain core skill sets should always be represented within the team, whether by a single individual or across distinct roles.

The first is the *ML Engineer,* who builds infrastructure hosting the pipeline. This involves developing interfaces that connect existing databases and operations systems. Application Programming Interfaces (APIs) are a key technology for this, enabling different software systems to communicate and interact, thus allowing automated dataflows into and out of the ML pipeline. ML Engineers ensure scalability, reliability, and automation, and migrate models into code suitable for production environments.

The next role is the *Data Scientist,* who specialises in the statistical modelling of ML algorithms. The Data Scientist explores and processes available data, selects appropriate ML algorithms, and analyses model performance to assess how well it generalises beyond the training data.

Finally, *Domain Specialists* play a critical role in defining contextually model inputs, validating outputs for realism and operational value, providing real-world context, and identifying risks or constraints. Their subject matter expertise often provides a valuable head start for the modelling phase.

Military personnel frequently serve as Domain Specialists, offering deep understanding of complex operational problem sets that academic or industry-based computer scientists may not fully grasp. It is a crucial role; therefore, individuals wishing to contribute to an ML pipeline but lacking technical expertise might consider volunteering as Domain Specialists.

3.2.2 Information Technology Backbone

The ML pipeline requires a solid IT backbone, including (but not limited to) a programming environment, sufficient data, robust computational resources, and monitoring tools. Storing, transporting, and processing large-scale data demands advanced hardware and software solutions. Further details on infrastructure requirements for ML model deployment are provided in Chapter 4, Section 4.4 – *Deployment and Monitoring*.

The programming environment typically relies on the Python coding language and its extensive *libraries,* which are associated modules built in Python. Common Python libraries that the ML Engineer should incorporate are shown below in Figure 3.1.[1] Many of the concepts associated with these libraries will be discussed throughout the next two chapters.

For ML tasks, historical data are used to train and test machine learning models, alongside a plan to ingest live data once the model enters production. Machine learning performs most effectively when the data used for training are similar to those the final product will encounter. For example, it would be difficult to derive meaningful insights if a model were trained on 1990s Chicago O'Hare International Airport air traffic and then applied to Bagram Airfield, Afghanistan in 2010. In this instance, while some *features,* or variables, might be similar, the overall context and datasets would differ significantly.

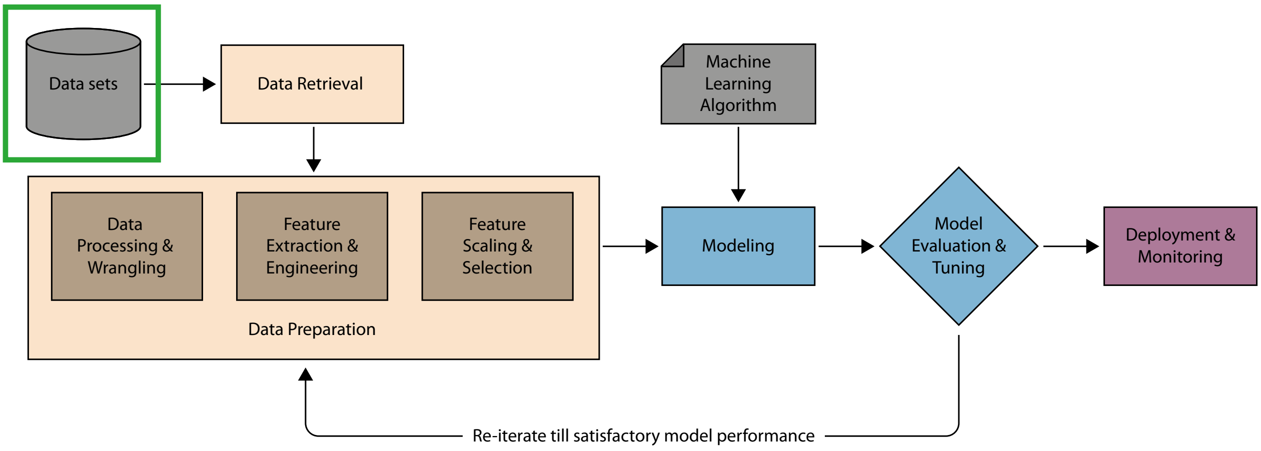
3.3 Understanding the Task and Data

When embarking on any ML endeavour, it is essential to thoroughly understand both the task and the data at hand. Although data are ubiquitous, identifying the right data to work with can be challenging; thus,

Python Libraries	Used for
NumPy	Mathematical functions that support complex matrices and arrays.
Pandas	Analysing and manipulating tabular datasets.
Matplotlib & Seaborn	Data analysis through visualisation.
Scikit-Learn	Traditional machine learning tasks; contains various classification, regressions, and clustering algorithms, and more.
SciPy	Scientific and mathematical problems.
TensorFlow	Deep learning in production due to scalability and high performance.
PyTorch	Deep learning in research due to dynamic graphs and easy debugging.
XGBoost	Boosted decision tree modelling optimised for accuracy.
LightGBM	Boosted decision tree modelling optimised for speed.

© JAPCC

Figure 3.1: *Examples of widely used Python libraries for machine learning. All these libraries are available to download today, without any payment or registration required.*


Data sets
Data Retrieval
Machine Learning Algorithm
Data Processing & Wrangling
Feature Extraction & Engineering
Feature Scaling & Selection
Data Preparation
Modeling
Model Evaluation & Tuning
Deployment & Monitoring
Re-iterate till satisfactory model performance

© Springer Nature BV, 2017, reproduced with permission.

understanding their availability and limitations is a prerequisite. Accordingly, a series of clarifications and guiding questions must be addressed before any data retrieval or preparation can begin. These are outlined in the following two sections.

3.3.1 Task Definition

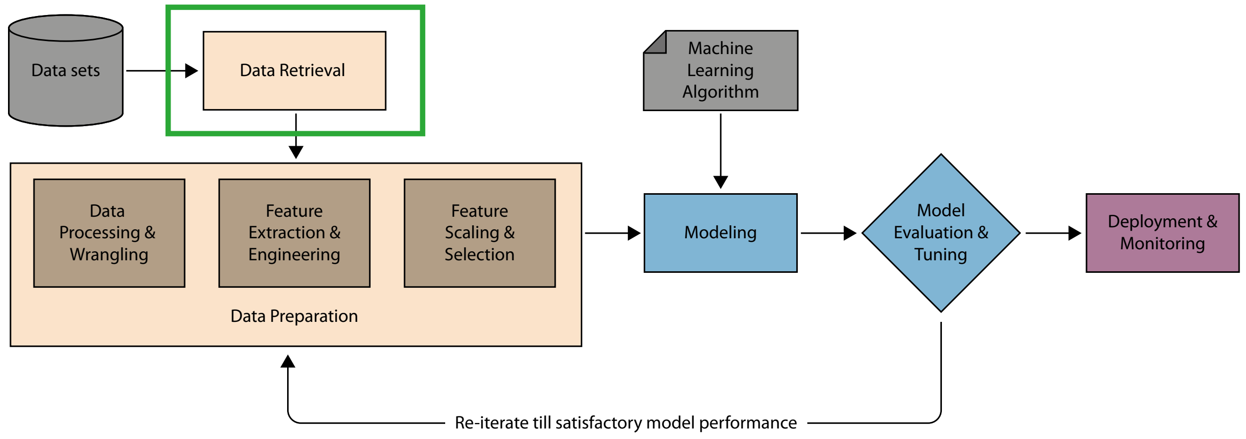
First, the team should determine whether AI/ML are indeed the appropriate tools for the problem. Often, the simplest solution that achieves the objective is the best one; introducing unnecessary complexity can create avoidable challenges. ML should be employed only when sufficient data are available, when outcomes cannot be predicted through human-made algorithms, and when meaningful predictions can be derived by identifying patterns.

Next, it is important to define the problem by specifying the type of ML prediction required. Do stakeholders seek classification, where a model categorises data into pre-labelled groups, or regression, which predicts numeric outputs from continuous inputs? Alternatively, unsupervised methods such as clustering, association analysis, and anomaly detection may be used, allowing the model to identify hidden patterns that Data Scientists and Domain Specialists can then explore further. Examples of these task types are below in Figure 3.2.

Task Type	Example
Classification	Identifying objects from satellite imagery as tanks, automobiles, or surface-to-air missile (SAM) launchers.
Regression	Determining stand-off weapon accuracy from a number of numerical factors related to the environment, range, targeting information, adversary defences, etc.
Clustering	Discovering patterns among maritime vessels using AIS data without predefined categories.
Association Analysis	Finding specific configurations and dependencies of large-scale IT networks that lead to more failures.
Anomaly Detection	Revealing unusual electromagnetic interference from a massive collection of radiofrequency transmissions.

© JAPCC

Figure 3.2: *Five classic types of machine learning problem sets and examples. Note that there are other more nuanced and advanced task types too, although these cover many use-cases.*



Data sets
Data Retrieval
Machine Learning Algorithm
Data Processing & Wrangling
Feature Extraction & Engineering
Feature Scaling & Selection
Data Preparation
Modeling
Model Evaluation & Tuning
Deployment & Monitoring
Re-iterate till satisfactory model performance

© Springer Nature BV, 2017, reproduced with permission.

Third, what performance metrics will define success? What level of error is acceptable? Are false-positives, where too many signals are detected, more tolerable than false-negatives, where critical signals may be missed? The Data Scientist and Domain Specialist will assess model performance through various statistical measures discussed in the *Model Evaluation and Tuning* section of the next chapter. However, it is vital to establish an initial understanding early on, as that will determine which outcomes to optimise for and guide decisions regarding data, features, algorithms, and evaluation criteria.

3.3.2 Data Orientation

The saying 'junk in, junk out' aptly applies to data within a ML pipeline. A model's quality is inherently limited by the data on which it is trained and tested. Therefore, the ML Engineer, Data Scientist, and Domain Specialist must all consider where the data originate, their format, and their structure. Are the data structured in clean tables captured from real-world databases? Do they derive from free-text intelligence reports in multiple languages, or a combination of both? Do they reflect what a future operational environment might look like? These questions critically inform the design of the ML pipeline.

For example, if a model is designed to predict incursions into NATO airspace, a supervised model cannot identify the incursions unless the training data include such instances (whether simulated or real). Without these, the model has no patterns from which to learn. These instances must therefore be clearly marked, a process known as labelling, so that the model understands what it is intended to predict.

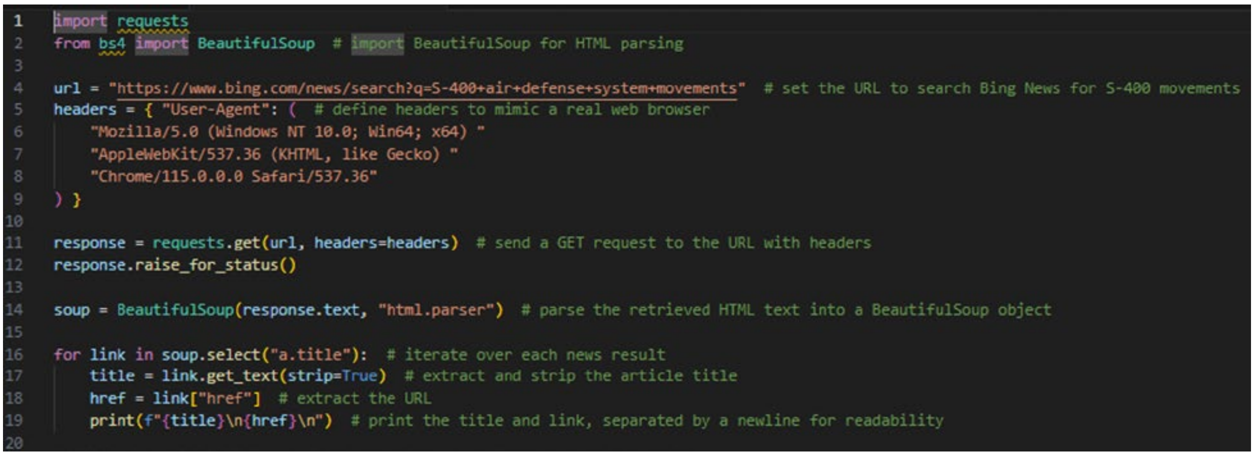
Labelling can be a time-intensive, manual process that almost anyone familiar with the problem can assist with. For instance, within the US military, an entire ISR Wing once sought volunteers to label satellite imagery; airmen from diverse backgrounds supported AI training by identifying objects within images. A commercial parallel is Google's reCAPTCHA system, which prompts users to select all fire hydrants or buses before accessing a web page, hypothetically training Google's AI models in the process.

All these considerations will determine the data retrieval methods and level of preparation required in the subsequent stages of the ML pipeline.

3.4 Data Retrieval

Identifying data-sourcing options is the first step in data retrieval. Many operational applications run with background databases that store large amounts of information. This is called tabular data, and it is an ideal candidate for integration into the ML pipeline due to the relative ease of transfer. However, structured and well-organised datasets are not always available; in these cases, Domain Specialists, Data Scientists, and ML Engineers must collaborate to identify alternative sources.

Logs are often valuable sources of operational data. They are typically recorded continuously and follow a



```python
import requests
from bs4 import BeautifulSoup  # import BeautifulSoup for HTML parsing

url = "https://www.bing.com/news/search?q=S-400+air+defense+system+movements"  # set the URL to search Bing News for S-400 movements
headers = { "User-Agent": (  # define headers to mimic a real web browser
    "Mozilla/5.0 (Windows NT 10.0; Win64; x64) "
    "AppleWebKit/537.36 (KHTML, like Gecko) "
    "Chrome/115.0.0.0 Safari/537.36"
) }

response = requests.get(url, headers=headers)  # send a GET request to the URL with headers
response.raise_for_status()

soup = BeautifulSoup(response.text, "html.parser")  # parse the retrieved HTML text into a BeautifulSoup object

for link in soup.select("a.title"):  # iterate over each news result
    title = link.get_text(strip=True)  # extract and strip the article title
    href = link["href"]  # extract the URL
    print(f"{title}\n{href}\n")  # print the title and link, separated by a newline for readability

```

© JAPCC

Precursor Python code for a tool that scrapes search engine results for headlines related to S-400 air defence mobilisation and saves them for subsequent trend- and meta-analysis.

consistent structure, although some processing may be needed to standardise formats. In a military context, examples include radar detection histories, maintenance records, IT system alerts, training logs, medical data, cargo manifests, fuel consumption levels, flight times, and more. In the absence of internal databases or logging systems, Domain Specialists and ML Engineers can work together to deploy sensors on systems or within networks to begin capturing data in a consistent manner.

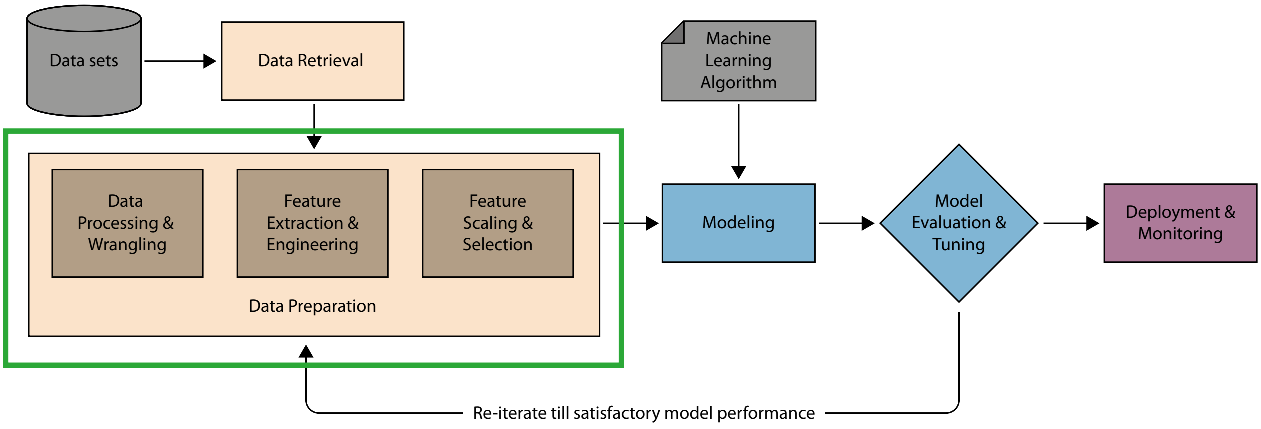
Internal data retrieval should be prioritised, as such data are more likely to reflect real-world conditions. If internal data is unavailable or insufficient, however, the team may need to gather external data from third-party sources. This often can be achieved by purchasing data from vendors or by using export tools provided by data repositories and services. In some cases, an ML Engineer might even design a web scraper to automatically collect relevant information from online sources, similar to how search engines index the internet. Within the military context, however, structured, secure, and consistent data are often difficult to obtain due to classification and operations security (OPSEC) constraints. Nevertheless, this step provides the essential foundation for a scalable and sustainable pipeline.

In supervised learning, labels designate the ground truth values that the model seeks to predict. For instance, in the classification example in Figure 3.2, labels are applied to imagery objects to indicate their true identity, such as a tank, automobile, or surface-to-air missile (SAM) launcher. In the regression example, labelling provides the real-world maximum stand-off weapon accuracy represented in the data. Pursuing supervised learning models is possible only when labelled data exist.

If labels are missing, Domain Specialist must either manually assign them or devise a scalable method to generate them in bulk. This introduces a trade-off between speed and scale on one hand, and accuracy and reliability on the other. Otherwise, only unsupervised learning techniques are feasible, where models operate without explicit guidance on what they are seeking.

During data retrieval, Domain Specialists should inform the ML Engineers of any risk that the data structure might change from that used to train the model. For example, if a new aircraft variant becomes operational and introduces new data features, the ML Engineer must develop a method to integrate this evolving information into the pipeline. The Data Scientist must then assess how the new or unexpected data influence the model's statistical performance.

Furthermore, it is critical to verify that no missing, erroneous, or corrupted data enter the ML pipeline, both during development and after deployment. Finally, because data often traverse multiple systems, strict handling protocols must be observed to prevent


Data sets
Data Retrieval
Machine Learning Algorithm
Data Processing & Wrangling
Feature Extraction & Engineering
Feature Scaling & Selection
Data Preparation
Modeling
Model Evaluation & Tuning
Deployment & Monitoring
Re-iterate till satisfactory model performance

© Springer Nature BV, 2017, reproduced with permission.

spillage between classification enclaves. Within NATO, each organisation and nation may have its own handling requirements that must be respected. In an ideal scenario, all 32 member nations would pool their data together to generate powerful insights; however, such collaboration remains unrealistic due to persistent security and releasability constraints.

3.5 Data Preparation

Once all relevant data are retrieved, the next stage involves comprehensive data preparation. This process is analogous to conducting Mission Analysis in the Operations Planning Process; if rushed, the entire pipeline is likely to fail. Data may enter the pipeline from multiple sources and in diverse formats; therefore, this stage focuses on transforming the raw inputs into a clean, unified dataset suitable for modelling.

Subordinate steps within this phase include data processing and wrangling, feature extraction and engineering, and feature scaling and selection. From the outset, it is essential to determine which features will serve as inputs to the model, and, where supervised learning is used, which feature is the target variable the model will attempt to predict.

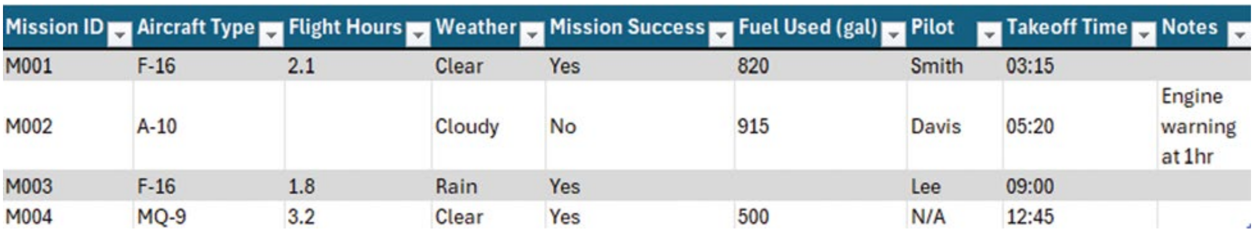
3.5.1 Data Processing and Wrangling

The first step in data preparation is processing and wrangling, transforming the raw data retrieved from one or more sources into a clean, usable format for the AI model. The Domain Specialist and Data Scientist must decide how to clean and standardise the data, as they will likely encounter missing values, outliers, and inconsistent formatting. Such inconsistencies may prevent an ML model from running or reduce its reliability.

3.5.1.1 Data Removal and Imputation

Two approaches are commonly used to manage missing values: removal or imputation. If the team opts to remove missing data, they may either delete individual datapoints containing missing values or remove entire features (attributes). For example, in a notional flight mission dataset (on the next page), each row might represent a mission, while each column represents a feature or attribute. If only a small fraction of the dataset contains missing entries, those datapoints can simply be dropped without significantly affecting the model. However, if a feature has many missing entries and is not critical to the analysis, it may be preferable to drop that feature altogether to retain more usable data.

Alternatively, *imputation* replaces missing values with substitute data. The Data Scientist and Domain Specialist must determine the most appropriate method of imputation, such as using the mean, median, or mode of the available data, randomised values, or more advanced techniques that involve further modeling. Regardless of the approach, it is inadvisable to ignore missing values, as the model may fail to execute or perform poorly. Only certain machine learning algorithms, typically decision-tree based models, can tolerate missing data natively.



Mission ID	Aircraft Type	Flight Hours	Weather	Mission Success	Fuel Used (gal)	Pilot	Takeoff Time	Notes
M001	F-16	2.1	Clear	Yes	820	Smith	03:15	
M002	A-10		Cloudy	No	915	Davis	05:20	Engine warning at 1hr
M003	F-16	1.8	Rain	Yes		Lee	09:00	
M004	MQ-9	3.2	Clear	Yes	500	N/A	12:45	

© JAPCC

A small, notional dataset of flight missions used to illustrate concepts within Section 3.5. Real-world AI/ML problems may involve hundreds of thousands of datapoints or more.

Outliers, or datapoints that deviate significantly from the norm, should also be assessed. The team must decide whether to remove them to prevent biasing the model or retain them if they present meaningful anomalies relevant to the problem being addressed.

3.5.1.2 Data Cleaning

Beyond addressing missing values, the formatting of data must also be standardised to ensure models operate correctly. For instance, ML algorithms cannot interpret contextual differences between measurements recorded in metres and yards, so consistent unit conversion is necessary. Likewise, certain features may need to be reformatted, for example, converting integer data types (which cannot contain decimals) into floating-point numbers (which can), without altering the underlying values. The same applies to date and time formats. While a human can easily recognise that 111293, 11.12.1993, and 12/11/1993 are the same date, an algorithm lacks the contextual understanding. Consistency in data formats across all features is therefore essential.

3.5.1.3 Data Encoding

For categorical data, such as the *Aircraft Type* feature in the notional flight dataset above, it is often useful to *one-hot encode* the values so they align with the rest of the numerical data. This process creates a new feature (column) for each category (e.g. *F-16, A-10,* etc.) and assigns a binary value: *1* if true, are *0* if false.

In the example dataset, it may be beneficial to one-hot encode the *Weather, Pilot,* and *Notes* features. By the end of the process, the dataset may contain many more features than it started with. If one-hot encoding produces an excessive number of variables (for example, if every NATO-aligned aircraft type were represented), alternate encoding methods, such as target, frequency, or label encoding may be more suitable.[2]

Another approach is ordinal encoding, which assigns a ranked integer value to each category (e.g. *A-10 = 1, F-16 = 2, MQ-9 = 3*). However, for this example, this could imply a hierarchy that does not exist.[3] Determining the 'best' aircraft is subjective and unsuitable for algorithmic interpretation. Ordinal encoding should only be applied when a true order exists, such as classifying aircraft by generation (4th gen = 4, 5th gen = 5, 6th gen = 6).

3.5.1.4 Class Imbalances

Another key consideration during this step is to resolve class imbalances for the target variable that has been labelled beforehand by the Domain Specialist. In the example dataset above, the task may be to figure out which combination of factors increases the chance of an engine warning, which is labelled within the *Notes* feature. This *Notes* data should first be one-hot encoded into a new feature, *Engine Warning*, where the value is labelled *0* when there was no warning and *1* when there was.

As it stands, three out of four, or 75 %, of the current data has no engine warning. If the algorithm processed the data as-is, it could model its predictions off the simple logic that there will never be an engine warning, and it would be 75 % accurate. While 75 % is a theoretical example chosen to illustrate a point, a more realistic case could be predicting malicious emails, which may only make up 0.01 % of total emails received. A model could predict with 99.99% accuracy that all emails are safe, but the underlying logic would make the model worthless.

To handle the dilemma, some models can assign more weight to the less common outcomes. If the algorithm cannot do so on its own, class balancing is required. This equalises the number of datapoints in the majority class (the datapoints without engine warnings) and the minority class (the datapoints with engine warnings).

If the Data Scientist assesses that there is sufficient data to do so, the first approach could be to undersample the majority class. That entails randomly removing datapoints from the majority class until the dataset has approximately a 50/50 balance of datapoints with and without engine warnings.

If that approach would remove too much data, it is also possible to cautiously oversample the minority class. This can be done by duplicating existing datapoints within the minority class, or by creating synthetic datapoints from scratch through techniques such as the *Synthetic Minority Over-sampling Technique (SMOTE)*. SMOTE, available in various Python libraries, generates new datapoints that are close enough to the existing ones that it should not disrupt the model's accuracy. However, it is always preferable to rely on original data when possible.

3.5.1.5 Exploratory Data Analysis

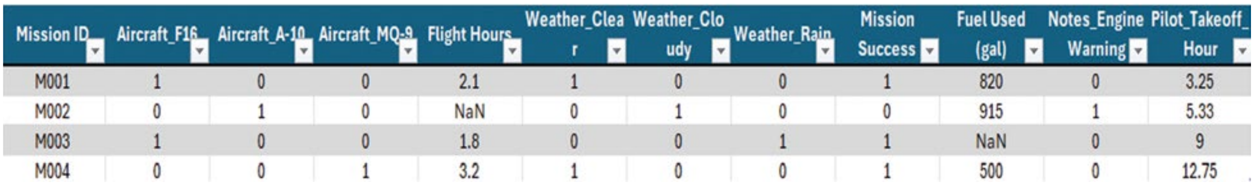
At this point, Exploratory Data Analysis (EDA) can be implemented to visualise distributions, trends, and correlations. It may be straightforward to detect any issues, outliers, or patterns in the small, notional flight mission data above, but if there are millions of datapoints and hundreds of features, it would be too much for the human eye (or even a desktop Microsoft Excel workbook).

EDA tools like *matplotlib* and *seaborn* can be used to examine plots, histograms, heatmaps, metadata, and graphs to uncover insights into the data that may be worth exploring further before proceeding. EDA is not the stage to confirm hypotheses, but rather to uncover unforeseen relationships within the data before proceeding. These visualisations and statistical reports can help inform the next step of data preparation, feature extraction, and feature engineering.

3.5.2 Feature Extraction and Engineering

The Domain Specialist will have an active role in feature engineering due to their deep knowledge of the subject matter at hand. Creativity at this stage can reduce unnecessary tweaking and re-running of models later, saving time and resources. One basic example from above is if the Domain Specialist notices during EDA that there is multicollinearity, or a strong correlation, between the *Fuel Used* and *Flight Hours* features from the notional dataset. This is reasonable, as flights that last longer typically consume more fuel. The Domain Specialist should consult with the Data Scientist to decide whether to remove one of these highly correlated features, as the model could overemphasise one phenomenon if it is represented twice.

It also may be useful to extract the data to create new features altogether. For example, *Fuel Used* could be misleading when comparing a small fighter to a heavy tanker. Developing a new feature that accounts for fuel usage relative to the aircraft type, flight duration, or another measure could provide a more meaningful metric. This new feature could be derived by dividing *Fuel Used* by *Flight Hours* to create a fuel efficiency measure.



Mission ID	Aircraft_F16	Aircraft_A-10	Aircraft_MQ-9	Flight Hours	Weather_Clear	Weather_Cloudy	Weather_Rain	Mission Success	Fuel Used (gal)	Notes_Engine Warning	Pilot_Takeoff_Hour
M001	1	0	0	2.1	1	0	0	1	820	0	3.25
M002	0	1	0	NaN	0	1	0	0	915	1	5.33
M003	1	0	0	1.8	0	0	1	1	NaN	0	9
M004	0	0	1	3.2	1	0	0	1	500	0	12.75

© JAPCC

The same dataset after one-hot encoding the 'Aircraft Type' and 'Notes' features. Categorical data is converted into new binary columns (0 or 1) to be mathematically processed by a model.

Furthermore, *Mission ID* is helpful for humans to catalogue each flight, but it is rather meaningless for a model to process without additional context. If the Domain Specialist can associate each *Mission ID* with a type of sortie (e.g. air-to-air engagement, close air support, counter-A2/AD) and group them accordingly, that could help the model detect relevant patterns. Creative, human-devised features at this stage can significantly improve the model's predictive performance. Only the most advanced deep learning models can handle this step automatically, so it is best to complete feature engineering before modelling.

3.5.3 Feature Scaling and Selection

Many ML algorithms are sensitive to differences in the range of values when comparing features, such as how *Flight Hours* may be in the tens while *Fuel Used* is in the hundreds. It can be difficult for models to interpret these differences, so scaling is often used to reconcile them.

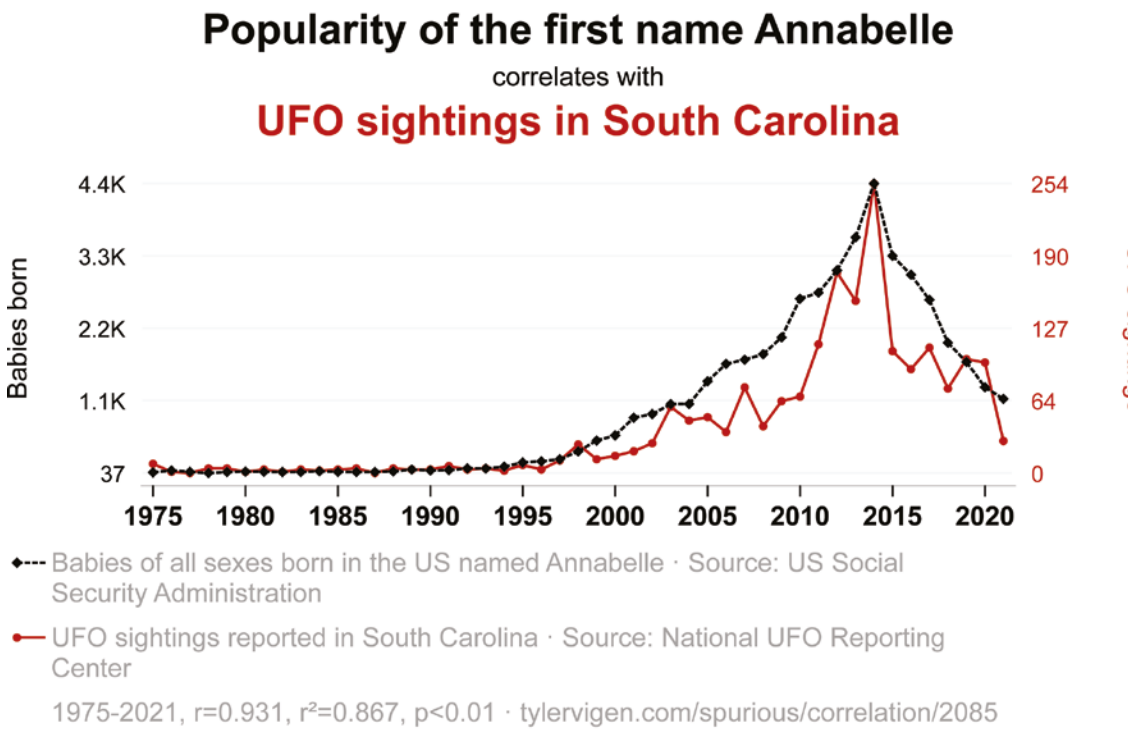
Tools within Python libraries such as *StandardScaler* and *MinMaxScaler*, use either standard deviations or minimum and maximum values to transform the values into standardised or normalised ranges. For instance, if all scaled features were normalised between 0 to 1, the largest values for *Fuel Used* and *Flight Hours* would both be 1. Scaling prevents the model from overemphasising features with larger magnitudes, such as *Fuel Used,* compared to smaller one like *Flight Hours.*

Important Note: Scaling (and other data modifications that alter values) should only be fitted to the *training data* after the dataset has been split to avoid look-ahead bias. More information on training, validation, and test splits will be discussed in Chapter 4, Section 4.2.1.

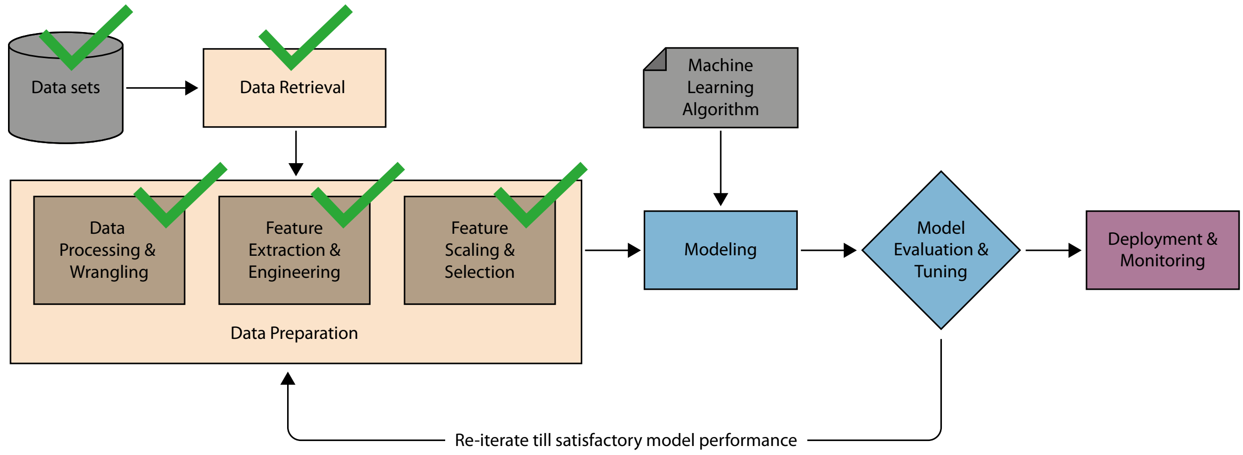
The final step before entering modelling is to select which features should be used. As mentioned in 3.5.2, removing a feature that is too correlated with another may be justified. Alternatively, a feature may simply be irrelevant and add unnecessary noise that the model must parse, such as the *Mission ID* feature. It is best to remove any irrelevant data that may bias the model towards unrealistic predictions. Coincidences in unnecessary data can lead to overfitting.

3.5.3.1 Feature Selection to Avoid Biases

Overfitting occurs when a model identifies meaningless patterns in random noise that appear significant



© tylervigen.com


Data sets
Data Retrieval
Machine Learning Algorithm
Data Processing & Wrangling
Feature Extraction & Engineering
Feature Scaling & Selection
Data Preparation
Modeling
Model Evaluation & Tuning
Deployment & Monitoring
Re-iterate till satisfactory model performance

© Springer Nature BV, 2017, reproduced with permission.

during training but do not generalise to new data. For example, if one were to compare the prevalence of the first name Annabelle in the United States with the number of unidentified flying objects seen in South Carolina between 1975 and 2020, the charts might appear similar. Intuitively, we know this is a coincidence, but with complex ML problem sets, such coincidences may not be as obvious.[4]

Another issue is look-ahead bias. If a model predicting engine warnings includes an *Early Return-to-Base* (RTB) feature, it could unfairly cheat by associating early RTBs with engine problems, information unavailable to planners before the flight. In this case, the model's predictions would appear accurate but would not reflect realistic conditions. **Both overfitting and look-ahead bias must be actively managed.** These and other mitigation strategies are discussed in the *Model Evaluation and Tuning section of Chapter 4.*

3.5.3.2 Automated Feature Selection

Automated feature selection tools can support the process by statistically evaluating which features are relevant. The *scikit-learn* Python library provides several useful methods.

Recursive Feature Elimination (RFE) runs a model multiple times, each time removing the least significant feature until only the most relevant remain.[5]

Principal Component Analysis (PCA) reduces features by identifying those that vary together (covariance) and combining them into new, composite features.[6] While PCA can simplify models and reduce processing time, it can also make them harder to interpret, which is an important consideration in military contexts where transparency and accountability are essential.

3.6 Conclusion

By this stage, data processing and standardisation should be complete, resulting in a clean, structured dataset for modelling. Missing or anomalous data should have been addressed, and relevant features extracted, engineered, scaled, and selected.

With this reliable dataset in place, the team can now move to the next major phase of the machine learning process: selecting, training, evaluating and deploying a model, which will be the focus of Chapter 4.

[1] DigitalOcean. (n.d.). Best Python libraries for machine learning in 2025. DigitalOcean. Retrieved 9 September 2025, from https://www.digitalocean.com/community/conceptual-articles/python-libraries-for-machine-learning.
[2] Burleigh, T. (19 September 2023). Encoding high cardinality features with 'embeddings'. Retrieved 9 September 2025, from https://tylerburleigh.com/blog/2023/09/19/.
[3] Stack Overflow. (2023). Ordinal encoding or one-hot encoding [Online forum post]. Retrieved 9 September, 2025, from https://stackoverflow.com/questions/69052776/ordinal-encoding-or-one-hot-encoding.
[4] Vigen, T. (2014). Spurious Correlations. Retrieved 9 September 2025, from https://www.tylervigen.com/spurious-correlations.
[5] scikit-learn developers. (n.d.). sklearn.feature_selection.RFE. In scikit-learn: Machine Learning in Python (v1.7.2). Retrieved 9 September 2025, from https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html.
[6] Priya C, B. (16 May 2023). Principal component analysis (PCA) with scikit-learn. KDnuggets. Retrieved 9 September 2025, from https://www.kdnuggets.com/2023/05/principal-component-analysis-pca-scikitlearn.html.

© Daniel – stock.adobe.com (Generated with AI)

Chapter 4

Machine Learning Pipeline (Part 2): Modelling, Evaluation, and Deployment

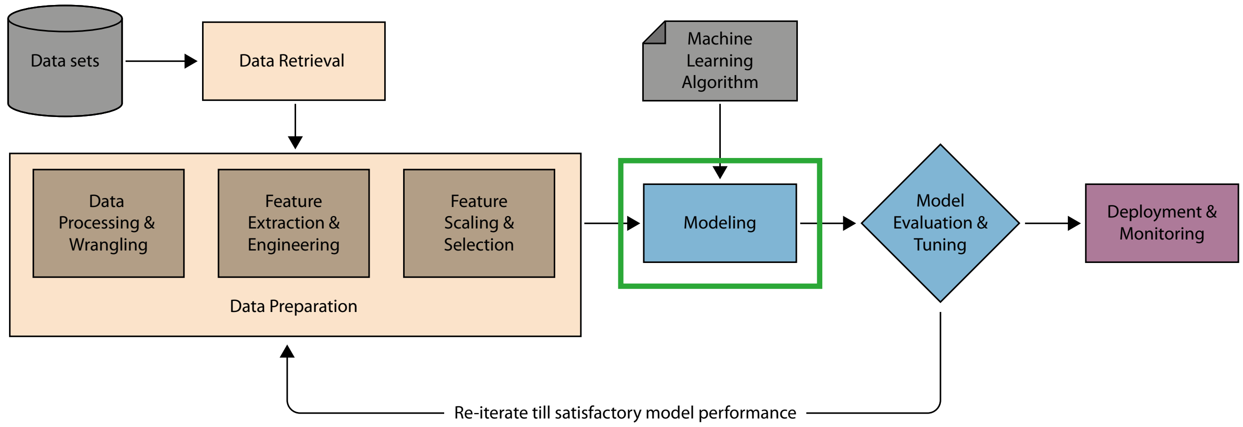
4.1 Chapter Overview

From this stage, the actual machine learning (ML) process begins. Before training the models, it is important to verify that all data has been prepared exactly as intended. Unlike working in an Excel spreadsheet, manipulating Pandas DataFrames in Python involves code-based operations that can introduce subtle human errors. Any missed step or unintended bias in the data can disrupt the training process, regardless of how advanced the algorithm is. Large Language Models (LLMs) can be helpful at this stage for reviewing code and identifying potential blind spots.

During modelling, evaluation, and tuning, the Data Scientist and Machine Learning Engineer may iteratively train several models and adjust their hyperparameters until one performs well enough for real-world deployment. The sections below outline each step in this process, from selecting and training models, to understanding *hyperparameters* (the human-defined settings that control how a model learns), and finally to deploying and monitoring the selected model in operation.

4.2 Modelling and Training

Once the data has been retrieved, analysed, and prepared, it should be divided into separate subsets for training, validation, and testing. The Data Scientist then selects and refines the most suitable algorithm for the task, coordinating with the Machine Learning Engineer to ensure the choice aligns with available computational resources and scalability requirements.


Data sets
Data Retrieval
Machine Learning Algorithm
Data Processing & Wrangling
Feature Extraction & Engineering
Feature Scaling & Selection
Data Preparation
Modeling
Model Evaluation & Tuning
Deployment & Monitoring
Re-iterate till satisfactory model performance

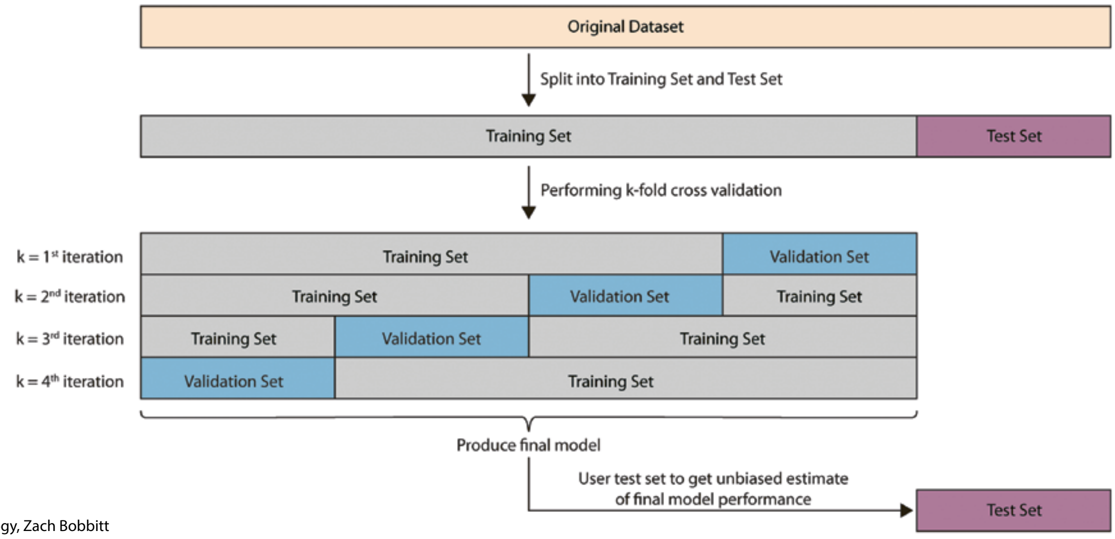
© Springer Nature BV, 2017, reproduced with permission.

4.2.1 Splitting Dataset into Train, Validation, Test Sets

Each portion of the dataset that is split will play a vital yet distinct role in the ML pipeline. The *training set* is the data that the algorithm learns from to produce a predictive model. The remaining data is used as the *test set* to evaluate the model's performance against unseen data. This approach reduces the risk of bias entering the pipeline by ensuring that the trained model can generalise to new information that it was not explicitly trained on.

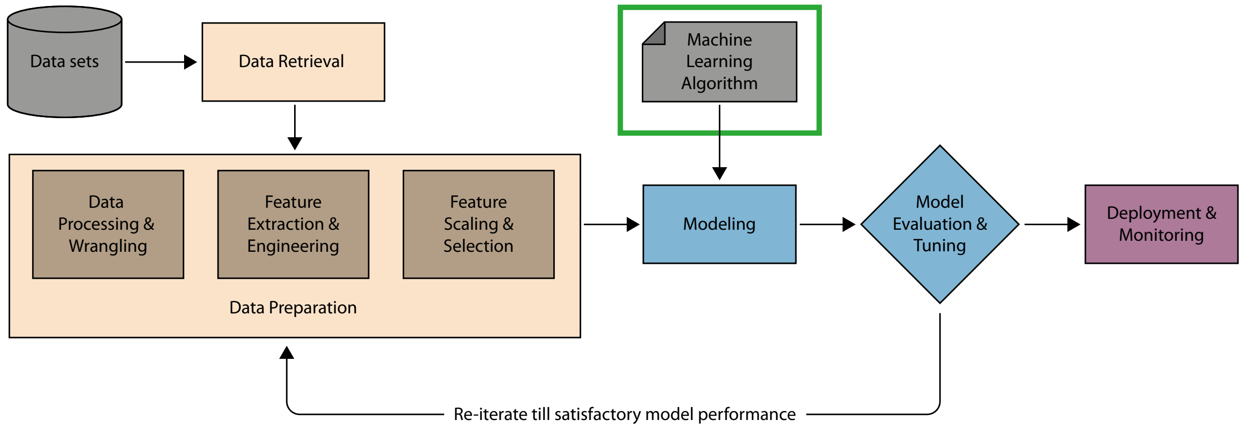
It is essential that no data from either set spills over into the other; otherwise, the model could cheat, leading to misleading performance metrics. The Data Scientist must decide on the proportion of data to split into each set, which is typically around 75–80 % for training sets and 20–25 % for test sets.

The Domain Specialist should be aware of how the data is divided, as the nature of the dataset can influence the splitting method. If the data is time-series based, the *TimeSeriesSplit* tool within the Python library *scikit-learn* can be used to establish splits that preserve chronological order and maintain cause-and-effect relationships. If the data does not have temporal importance, the *train_test_split* function (also within *scikit-learn*) can randomly assign data points across the full



© Statology, Zach Bobbitt

There is no one single best method for dividing data into training, test, and validation splits; however, the figure above illustrates a common training/test split with k-fold cross-validation. This specific approach should not be used if maintaining chronological order is important.[12]


Data sets
Data Retrieval
Machine Learning Algorithm
Data Processing & Wrangling
Feature Extraction & Engineering
Feature Scaling & Selection
Data Preparation
Modeling
Model Evaluation & Tuning
Deployment & Monitoring
Re-iterate till satisfactory model performance

© Springer Nature BV, 2017, reproduced with permission.

timeline while maintaining the desired proportions between training and test sets.

It is also common to include one more category: the *validation set*. Validation sets may be entirely separate from the training and test splits, or they may be subsets drawn from the training set when data is more limited. As the model undergoes each training iteration, called an *epoch, it* learns from its performance on the validation subset and adjusts its model parameters accordingly.

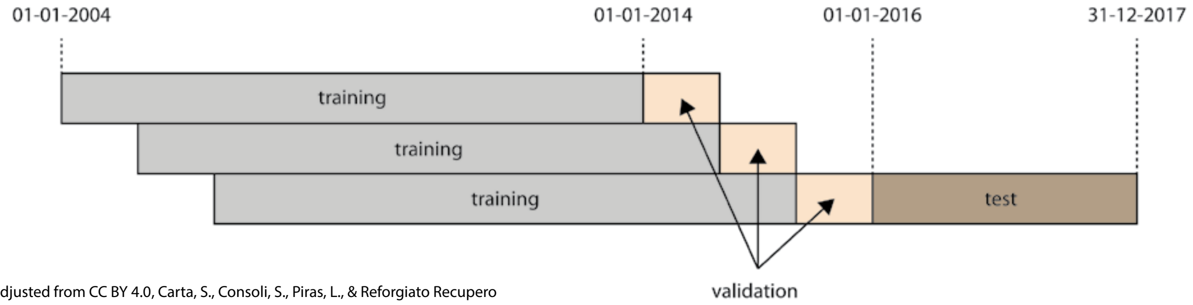
There are several ways to accomplish this, one being *k-fold cross-validation,* as shown in the visual on the previous page. In this example, there are four k-folds. Notice that for each fold, the validation data is simply a new subset within the original training set.

If the Data Scientist determines that the dataset is too small for k-fold cross-validation, *Leave-One-Out Cross-Validation* (LOOCV) is another approach. Here, each datapoint is used as the validation subset in turn, until the model has been validated against every sample individually. The overall model performance is then averaged across all iterations.[1] However, the number of iterations makes this approach computationally intensive.

Regardless of the splitting method, after each epoch, the model refines its parameters to improve performance in the next round of training. Once the model reaches a defined performance threshold, often set by human verdict, it is finally tested against the test set.

4.2.2 Machine Learning Algorithms

Figure 3.2 in Chapter 3 outlined the main types of machine learning problems that can be solved by their task type, such as classification, clustering, regression,



© Adjusted from CC BY 4.0, Carta, S., Consoli, S., Piras, L., & Reforgiato Recupero

To preserve chronological context within data splits, the 'walk-forward' approach can be used to assign validation subsets. In some cases, holdout time periods are placed between training and validation sets to act as firewalls, ensuring no data leakage occurs between them.[13]

association analysis, and anomaly detection. At this stage in the ML pipeline, it is time to assign an appropriate algorithm to the task at hand. Within each task type, there are a variety of ML algorithms that can be used.

While there are many options when deciding which models to use in a pipeline, it is prudent to choose the simplest model that can solve the task at hand. This reduces the chance of overfitting, when a model learns from the training data too closely and fails to generalise to new data.

Reference Figure 3.1 from Chapter 3 for common Python libraries that contain modelling tools discussed in this chapter.

4.2.2.1 Classification Algorithms

Classification, as mentioned in the previous chapter, is a supervised learning technique in which a model learns to assign data points to predefined categories, known as *labels*. The label that the model will predict is the *target variable*. These can be binary classifications (for example, routine computer network traffic versus malicious traffic) or multi-class (for example, classifying an object as a tank, automobile, or SAM). Even numerical data can be converted into classes by grouping them together in ranges, such as speed ranges of Mach 0–1, Mach 1–2, Mach 2–3, and so on.

To train a classifier, the model analyses patterns in the labelled training data and updates its parameters as it trains over each epoch. After each iteration, its performance is evaluated on validation sets to guide further tuning. Once it reaches a performance threshold, often pre-defined by the Data Scientist, it is tested against never-before-seen data from the test set.

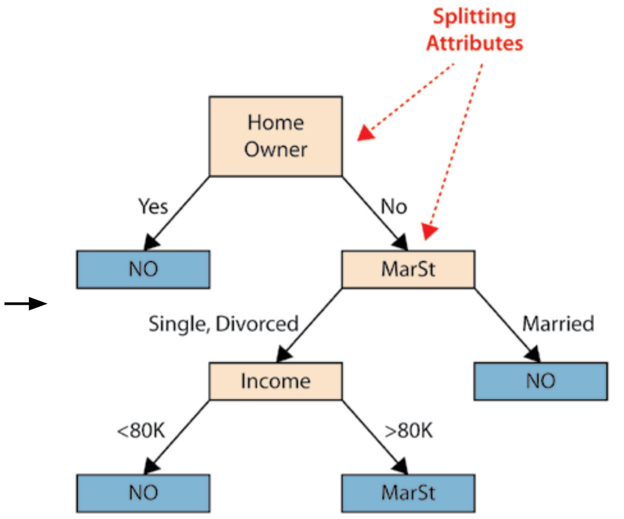
Classification algorithms may vary in how complex their rule sets are, their data requirements, and human interpretability of the final model. The Data Scientist needs to decide how complex the model should be. Too much complexity risks overfitting to coincidences within the dataset that may not generalise moving

Model Type	Considerations
Logistic Regression	Used for binary classification primarily.
Naïve Bayes	Simple model that can train quickly and does not tend to overfit.
K-Nearest Neighbour	A 'guilty by association' algorithm that categorises based on data points' proximity to other data points in the training set.
Decision Trees	Resembles a family tree in which each node, from root to leaf, represents a binary decision that a human can follow. Overly large and complex trees are prone to overfitting.
Random Forests	Combine multiple decision trees voting in parallel into one consolidated output.
Support Vector Machine	Boundaries are drawn between classes to distinguish them in a multi-dimensional plot. If there are more than two classes, it layers together multiple binary classifications.
Gradient Boosting	Layers together many simple predictors such as basic decision trees in sequence. LightGBM and XGBoost are examples.
Neural Networks	Often used with deep learning techniques for the most complex tasks. More detailed information about neurons and hidden layers can be found in Chapter 2.
Ensemble	Using multiple models, whether the same type or different types, in parallel or sequentially.

***Figure 4.1:** Some commonly used classification algorithm types.*[14]

© JAPCC

	categorical	categorical	continuous	class
ID	Home Owner	Marital Status	Annual Income	Default Borrower
1	Yes	Single	125 K	No
2	No	Married	100 K	No
3	No	Single	70 K	No
4	Yes	Married	120 K	No
5	No	Divorced	95 K	Yes
6	No	Married	60 K	No
7	Yes	Divorced	220 K	No
8	No	Single	85 K	Yes
9	No	Married	75 K	No
10	No	Single	90 K	Yes



Training Data

Model: Decision Tree

An example of a decision tree from the NSO Machine Learning and Artificial Intelligence for Cyber Applications course. 'Defaulted Borrower' is the target variable the model tries to predict.[15]

forward. Not enough complexity may miss crucial patterns. To find the balance, there are configurations that can be adjusted within the model's *hyperparameters*, the algorithm settings established before any training occurs.

4.2.2.2 Regression Algorithms

Regression modelling builds functions that ingest continuous, numerical features and predict continuous values. In other words, the model takes multiple independent variables to predict a dependent variable. Its success is often determined by a *goodness-of-fit* measure, also known as the *coefficient of determination (R^2)*. R^2 values range from 0 to 1, with 1 representing a perfect fit. Regression models are easier to interpret than other complex machine learning models, but that benefit comes with the trade-off of being suited for more routine tasks.

4.2.2.3 Clustering Algorithms

Clustering is an unsupervised technique where the model discovers natural groupings of data points without Domain Specialists labelling them beforehand. Imagine thousands of data points scattered around a two-dimensional plot. Clustering models would draw boundaries between various groupings to highlight relationships. While humans may be able to intuitively cluster small amounts of data portrayed in two-dimensional space, these models can analyse much larger datasets and are not limited to two dimensions.

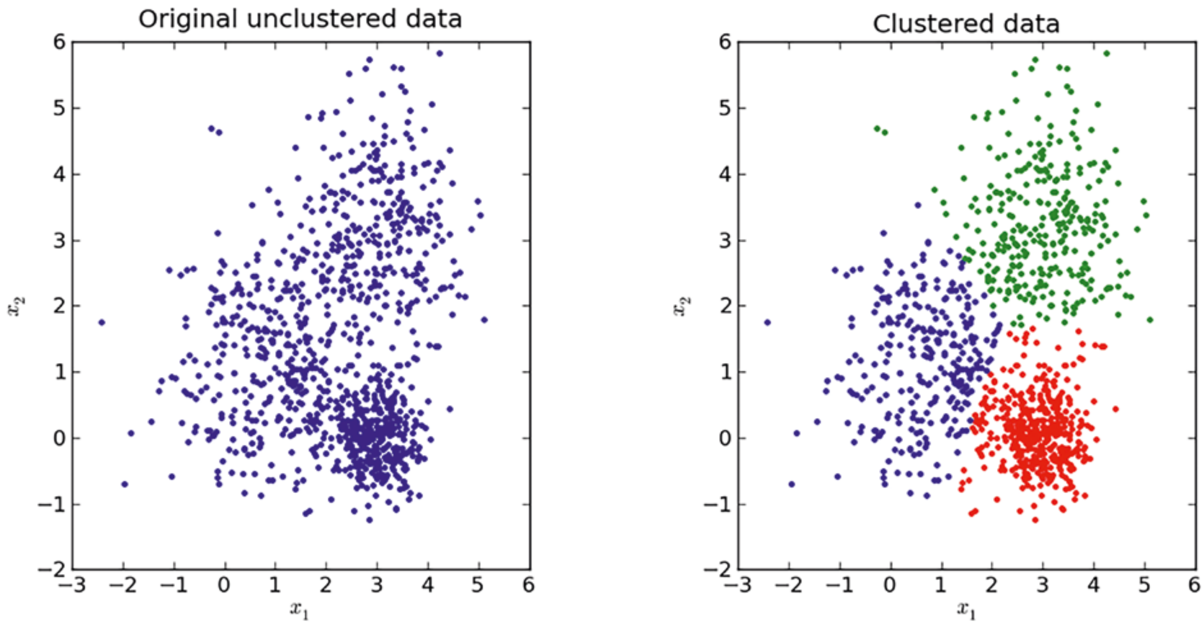
There are several algorithms that can decide how to draw those boundaries, based on densities, centroids, distributions, and hierarchies. If the Domain Specialist knows how many clusters to anticipate, they can provide that information to the algorithm. However, clustering can also operate without that input, albeit less predictably.

Clustering types	Algorithm(s)	Description
Centroid-based	K-Means	Creates clusters based on mean points within a cluster. Requires a predefined number of clusters and performs best when clusters are of similar size.
Density-based	DBSCAN, HDBSCAN	Clusters together dense regions. Does not require predefined cluster numbers but will not cluster outliers. Struggles with clusters of widely varying densities, or high-dimensional data.
Hierarchical	Agglomerative, BIRCH	Clusters through tree-like arrangements that have attributes at each node between branches.
Distribution-based	GMM	Assumes probabilistic distributions, such as Gaussian distribution (bell curve).

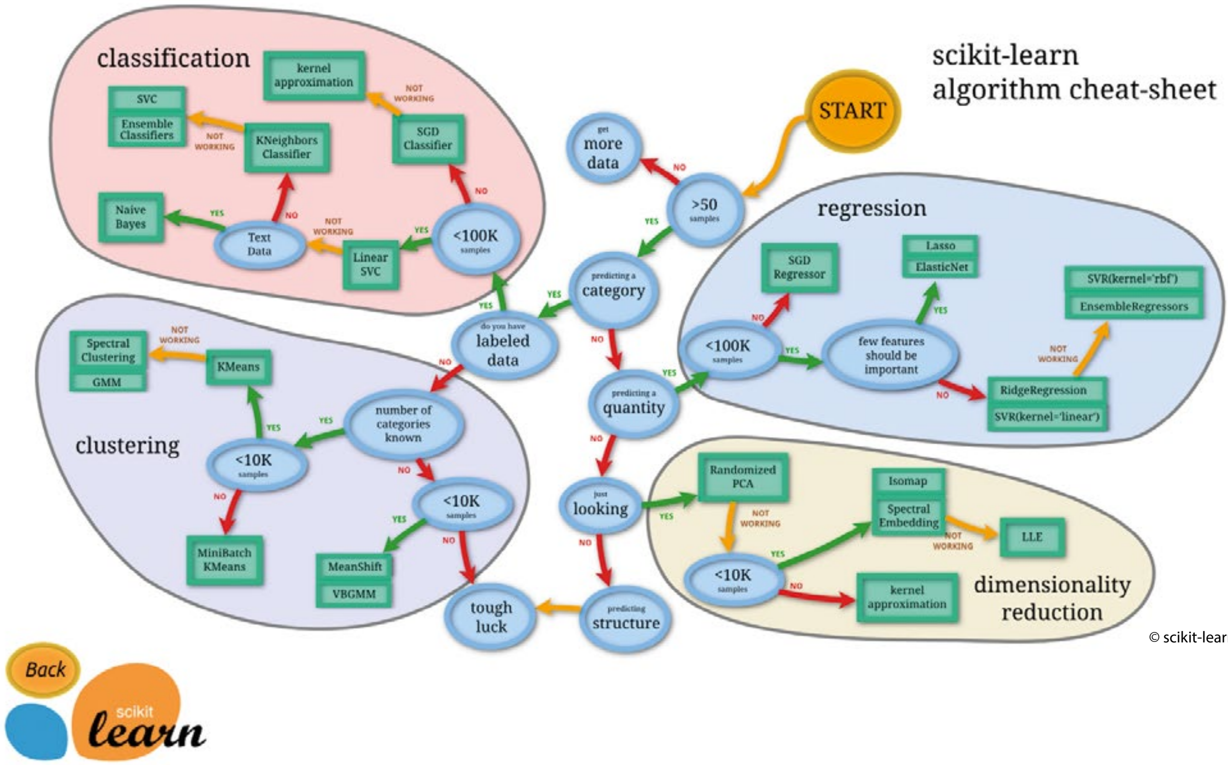
***Figure 4.2:** Some commonly used clustering types and associated algorithms.*[16]

4.2.2.4 Association Analysis Algorithms

Association Analysis identifies out hidden patterns and relationships within large datasets, such as when two events occur together and a third is also likely to occur. The *Apriori* algorithm, available in the *mlxtend* Python library, is often used to detect such patterns in machine learning contexts. It was originally designed to data mine through customer transactions to uncover 'frequent' pairings.[2] The Domain Specialist and Data Scientist determine what percentage 'frequent' means for their particular task. A frequency measure, known as *support*, could be set to 0.5 if two items occurring together 50% of the time meets their threshold. The algorithm analyses all possible combinations in large datasets, identifying relationships that exceed the predefined support level. Another measure, *confidence*, evaluates how likely one item is to occur given that



An example of the before and after of K-Means clustering.



© scikit-learn

Scikit-learn provides a flowchart to assist with algorithm selection, as well as extensive documentation and examples available online.[17]

another has already occurred. Together, support and confidence help reveal the strength and reliability of these associations.[3]

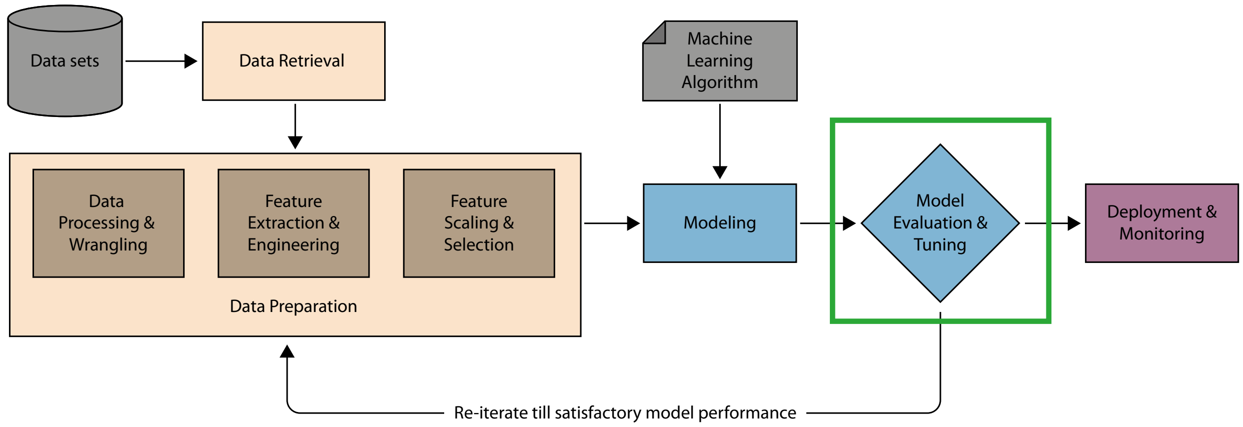
4.2.2.5 Anomaly Detection Algorithms

Anomaly Detection identifies abnormal data points that warrant further investigation. These anomalies might represent system malfunctions, computer network intrusions, satellites deviating from their intended orbits, atypical flight paths, or other irregular events. Anomaly detection leverages significant computing power to locate these 'needles in the haystack,' especially when could have important operational implications. The Domain Specialist helps define what constitutes normal behaviour, and what deviations to monitor. Working together, the Domain Specialist and Data Scientist can adjust the model's sensitivity so that it does not overwhelm users with too many alerts, while still capturing the anomalies that matter most.

4.2.2.6 Ensemble Methods Using Multiple Algorithms

Multiple models, whether the same or different types, can be combined within a single pipeline. They may operate in parallel, like voting systems where each model to a weighted vote to the final decision, or sequentially, like an assembly line where each model handles part of a broader problem. Both configurations are examples of *ensemble methods*, which can add useful robustness that often comes with one overly complex model.

Once the appropriate algorithms have been selected and configured, a simple press of the 'Enter' key can initiate training. Depending on model complexity, hyperparameter settings, data volume, and available computing resources, training can take anywhere from seconds to days. The output is a statistical model trained to make predictions based on the features it has learned.


Data sets
Data Retrieval
Machine Learning Algorithm
Data Processing & Wrangling
Feature Extraction & Engineering
Feature Scaling & Selection
Data Preparation
Modeling
Model Evaluation & Tuning
Deployment & Monitoring
Re-iterate till satisfactory model performance

© Springer Nature BV, 2017, reproduced with permission.

4.3 Model Evaluation & Tuning

The first attempt at modelling rarely produces the final version ready for deployment. Therefore, the Domain Specialist and Data Scientist must evaluate the model's performance using various metrics and adjust it accordingly. Tuning may involve revisiting data preparation, training a new model type, or adjusting the hyperparameter settings of the existing model.

4.3.1 Model Evaluation Metrics

After training, it is standard practice to generate performance metrics. For classification and anomaly detection tasks, common measures include *accuracy, recall, precision,* and *F1 score*. These metrics are derived from the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), which are often displayed together in a *confusion matrix,* such as the one below.

4.3.1.1 Accuracy Metric

Accuracy is calculated as the number of correct classifications divided by the total number classifications: (TP + TN) / (TP + TN + FP + FN).[4] The Domain Specialist should consider whether certain types of mistakes, such as false positives, are more costly than the others, such as false negatives, since accuracy does not account for these distinctions. As mentioned in Section 3.5.1.4, if an anomaly occurs only 0.1 % of the time but that 0.1 % is critical, accuracy alone is not a reliable measure. A model could achieve 99.9 % accuracy simply by always predicting that no anomaly occurs.

4.3.1.2 Recall Metric

Recall, or the *true positive rate*, measures the proportion of actual positives that were correctly identified. Its formula is TP / (TP + FN). Recall better addresses rare events occurring because it only gets credit for predicting the anomalies.[5] Recall is particularly effective for evaluating rare events because it only rewards the correct identification of positive cases, such as anomalies or failures.

Recall is best suited to situations where false negatives are more harmful than false positives. However, this comes at the cost of potentially increasing noise, as higher recall often leads to more false alarms. A related

	Model Predicted 'No'	Model Predicted 'Yes'
Actual 'No'	# of True Negatives (TN)	# of False Positives (FP)
Actual 'Yes'	# of False Negatives (FN)	# of True Positives (TP)

© JAPCC

***Figure 4.3:** The table above shows the structure of a confusion matrix. The number of TNs, FPs, FNs, and TPs are used to calculate evaluation metrics such as accuracy, recall, precision, and F1 score.*

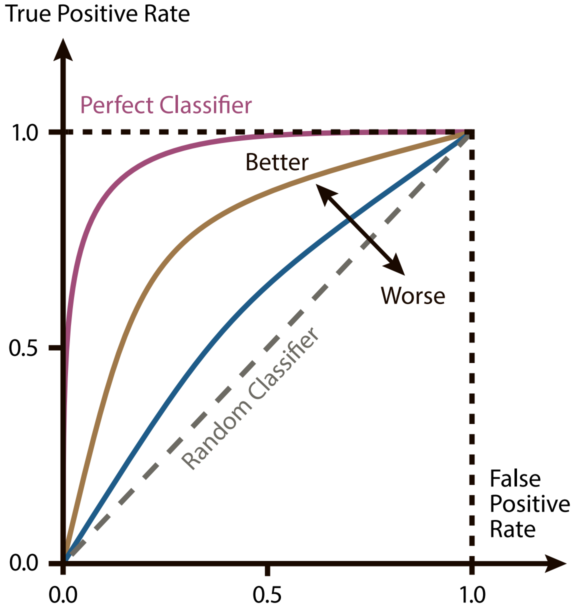
measure, the *false positive rate,* captures the rate of false alarms among true negatives and is calculated as FP / (FP + TN).

4.3.1.3 Precision Metric

Precision measures how many of the instances predicted as positive were actually correct. Its formula is TP / (TP + FP).[6] Precision serves as a safeguard against a model raising too many false alarms, since the metric decreases as the number of false positives increases.

4.3.1.4 F1 Score Metric

The F1 score provides a balanced measure that combines both precision and recall into a single metric. It is defined as 2 x (Precision x Recall) / (Precision + Recall). After simplification, the formula becomes 2TP / (2TP + FP + FN). The F1 score is especially useful when both false positives and false negatives carry significant consequences, providing a single, consolidated measure of model performance.



[M] © ProfDesigner/Shutterstock.com

Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) scores are visual tools that compare a model's recall (true positive rate) against its false positive rate. Plotting ROC curves and AUC values allows Data Scientists to quickly and visually compare the performance of multiple models.

4.3.1.5 Metrics for Regression Models and Interpretability

While the classification models use confusion-matrix-based metrics, regression models rely on different measures. In addition to the coefficient of determination (R^2), which has already been discussed, two important metrics are *mean squared error (MSE)* and *root mean squared error (RMSE).* MSE measures the average squared distance between the regression line and the actual data points, while RMSE is simply the square root of the MSE. Regardless of whether the Data Scientist and Domain Specialist are evaluating classification or regression models, it is important to decide which metrics are most relevant to the task before attempting to optimise model performance further.

Finally, interpretability tools such as *Shapley Additive Explanations (SHAP)* and *Local Interpretable Model-Agnostic Explanations (LIME) are* valuable for external stakeholders because they visualise model outputs in a way that is understandable to humans.[7] This interpretability helps the Domain Specialist and Data Scientist see how the model is making decisions, reducing its mysterious, black-box nature. These visualisations highlight how much each feature contributes to the model's predictions that are made. In doing so, they provide a more holistic view of model performance that can easily be communicated to non-technical audiences, including leadership.

4.3.2 Hyperparameter Tuning

Hyperparameter tuning is almost always required during model evaluation and tuning. Hyperparameters are top-level model settings that determine how the model learns. They differ from the parameters within the trained model because they are defined outside of the training process and are not influenced by the dataset.

Tuning hyperparameters can be done manually or automatically. In manual tuning, the Data Scientist uses experience and insights from earlier data preparation stages to explore various hyperparameter combinations. Automated approaches within the *sci-kit learn* library include *Grid Search, Random Search,* and *Bayesian Optimisation:*

- **Grid Search** exhaustively tests all possible hyperparameter combinations to identify the best one.
- **Random Search** tests a random subset of combinations, which is less resource-intensive but may miss the optimal configuration.
- **Bayesian Optimisation** learns from each trial and uses probabilistic models to more efficiently find the most promising hyperparameter combinations.[8]

Tweaking hyperparameters is crucial to finding the right balance in model complexity. A model that is too complex may overfit, meaning it becomes overly influenced by noise or coincidences in the training data that will not occur again in the live data. Conversely, a model that is too simple may underfit, failing to capture meaningful relationships within the data.

An underfit model typically performs poorly on both the training and validation/test sets. For example, a classification model might achieve accuracy no better than random guessing. In a deep learning task, the model might distinguish a tank from a civilian truck better than chance, but not with enough accuracy to be employed live without time-consuming human oversight.

By contrast, an overfit model performs very well on training data but poorly on unseen test data. Such a model cannot generalise beyond its training set and risks being unreliable in real-world use. Beware of those selling overfit models, as they are the AI equivalent of deceitful snake oil.

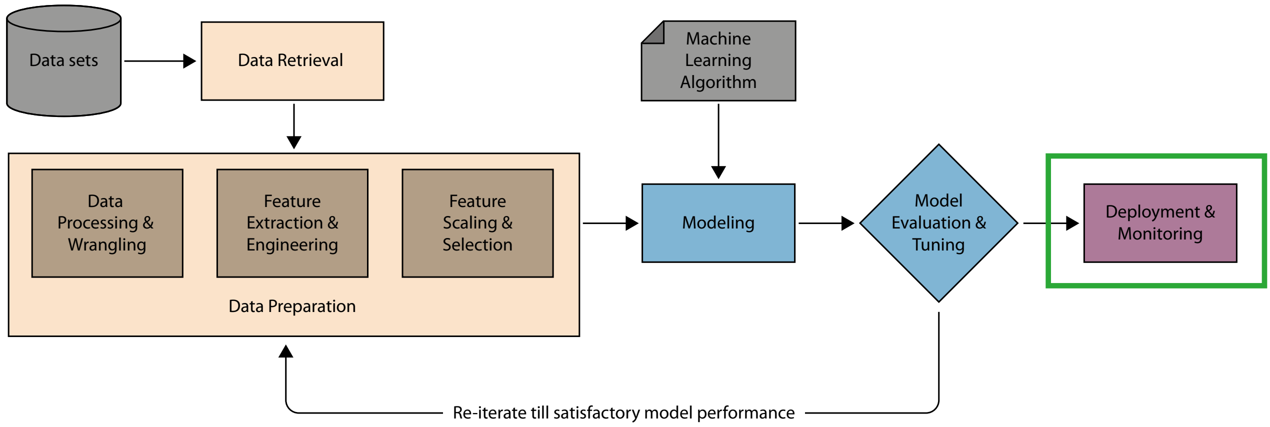
4.4 Deployment and Monitoring

At this stage, the ML Engineer needs to bring the model into a production environment so that it can make live predictions on new data. Storing and processing the vast amounts of data often required to generate predictions requires hardware storage and robust computational resources.

Hyperparameter	Description
Learning rate	Degree to which the model updates its parameters during training.
Neural network nodes	Number of neurons in a hidden layer. More nodes increase capacity but raise the risk of overfitting.
Neural network layers	Number of hidden layers. More layers enable the model to handle more complex tasks but make training more difficult.
Mini-batch size	Number of training examples processed before parameters are updated.
Epochs	Number of full passes through the training dataset.
Dropout	Randomly ignores certain nodes during training to prevent overreliance on specific ones, improving robustness.
L1 Regularisation	Sets some node weights to zero to eliminate their influence, helping the model generalise better.
L2 Regularisation	Penalises large weights shrinking them towards zero, providing a milder form of regularisation.
Early stopping	Stops training when performance on validation data no longer improves, reducing the risk of overfitting.

© JAPCC

Figure 4.4: *A sampling of hyperparameters, focusing on those used in neural networks due to their high degree of customisation.*[18]

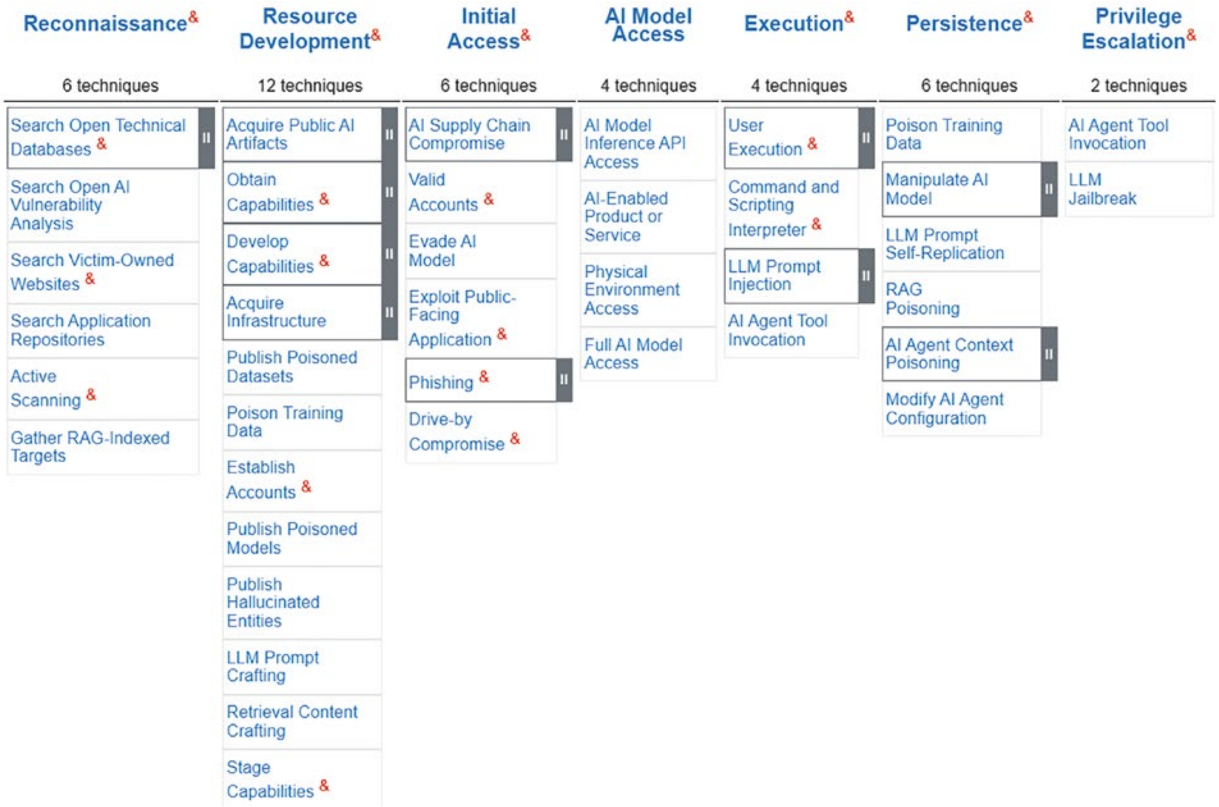

Data sets
Data Retrieval
Machine Learning Algorithm
Data Processing & Wrangling
Feature Extraction & Engineering
Feature Scaling & Selection
Data Preparation
Modeling
Model Evaluation & Tuning
Deployment & Monitoring
Re-iterate till satisfactory model performance

© Springer Nature BV, 2017, reproduced with permission.

4.4.1 Deployment Scalability

Traditional IT environments rely on Central Processing Units (CPUs), which can train models but are limited in handling large-scale computations efficiently. Graphics Processing Units (GPUs) and Tensor Processing Units (TPUs) have driven the recent growth in AI by enabling complex computations to run in parallel rather than sequentially.[9] Building data centres to house this type of hardware is a complex and costly undertaking, so cloud-based solutions for data storage and computation often provide a more practicable and scalable alternative.

A virtual environment is often created to simplify scaling and development. Within this environment, ML Engineers frequently use *containerisation* to package ML models and their dependencies in a modular, scalable, and repeatable way. Containerisation ensures consistency across deployments by bundling all necessary virtual components together, allowing for reliable migration to new environments.[10] Not all users (such as typical military personnel) can code, so ML Engineers must transition the final model into a user-friendly application. This application should be capable of automatically



Reconnaissance &	Resource Development &	Initial Access &	AI Model Access	Execution &	Persistence &	Privilege Escalation &
6 techniques	12 techniques	6 techniques	4 techniques	4 techniques	6 techniques	2 techniques
Search Open Technical Databases &	Acquire Public AI Artifacts	AI Supply Chain Compromise	AI Model Inference API Access	User Execution &	Poison Training Data	AI Agent Tool Invocation
Search Open AI Vulnerability Analysis	Obtain Capabilities &	Valid Accounts &	AI-Enabled Product or Service	Command and Scripting Interpreter &	Manipulate AI Model	LLM Jailbreak
Search Victim-Owned Websites &	Develop Capabilities &	Evade AI Model	Physical Environment Access	LLM Prompt Injection	LLM Prompt Self-Replication	
Search Application Repositories	Acquire Infrastructure	Exploit Public-Facing Application &	Full AI Model Access	AI Agent Tool Invocation	RAG Poisoning	
Active Scanning &	Publish Poisoned Datasets	Phishing &			AI Agent Context Poisoning	
Gather RAG-Indexed Targets	Poison Training Data	Drive-by Compromise &			Modify AI Agent Configuration	
	Establish Accounts &					
	Publish Poisoned Models					
	Publish Hallucinated Entities					
	LLM Prompt Crafting					
	Retrieval Content Crafting					
	Stage Capabilities &					

© The MITRE Corporation

ingesting new data as it becomes available and producing predictions without requiring extensive technical intervention.

4.4.2 Monitoring

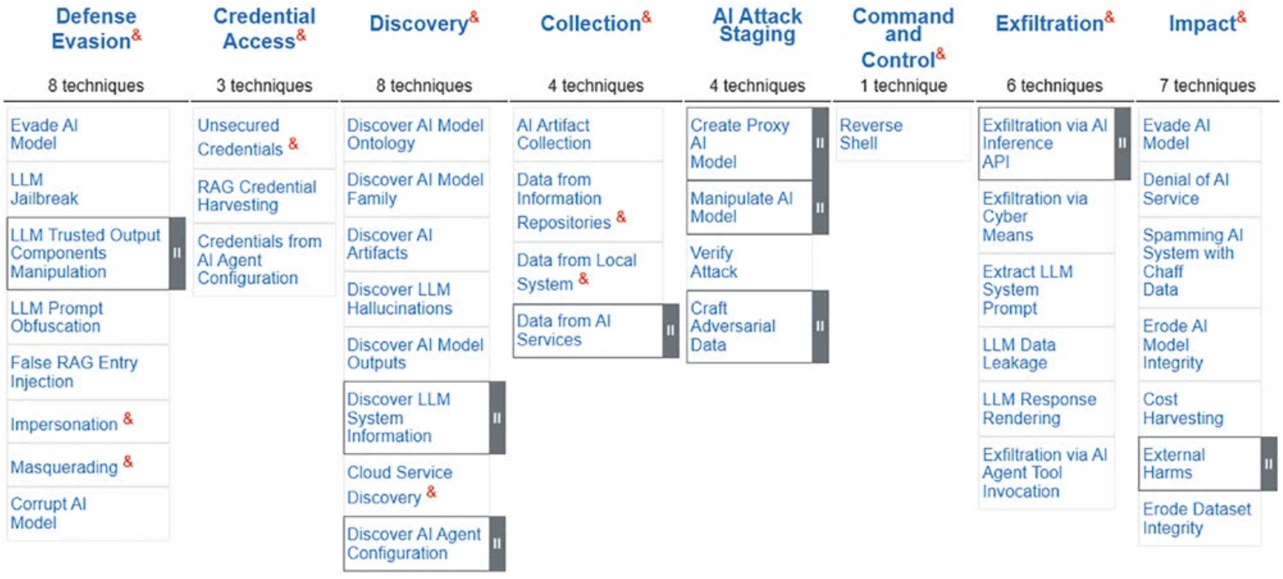
Once live, monitoring tools are essential to track and log model performance over time. These tools can monitor both the health of the IT systems for debugging purposes and the behaviour of the model itself for signs of performance degradation.

For example, if a model could characterise fourth generation aircraft with accuracy, it may face *data drift* as nations gradually adopt fifth generation aircraft that the model has not been exposed to as frequently during training. Once sixth generation aircraft become operational, there could be *concept drift,* which is a broader change in the patterns the model predicts from altogether. Either type of drift can degrade model performance and may require retraining.

All stakeholders must determine how frequently it is appropriate to retrain and redeploy models. It is recommended to establish checkpoints linked to decision criteria for whether to retrain or retire a model. In the military context, this may correspond to Commander's Critical Information Requirements (CCIRs) that trigger review when an operational ML model's performance falls below a predetermined threshold. The commander may then decide whether to accept the associated risk and continue using the model, retrain a new one, or transition to an alternative solution.

4.4.3 Cybersecurity

Finally, the ML Engineer must secure the entire pipeline, from the original datasets to deployment and monitoring. If an adversary were to compromise the integrity of the data or the model itself, it could create asymmetric effects, particularly since machine learning models may one day underpin critical operations. To mitigate such threats, many organisations have established Machine Learning Security Operations (MLSecOps), an emerging discipline within the broader field of AI and cybersecurity. CrowdStrike, a leading cybersecurity company, has identified several key risks that MLSecOps workflows must address by expanding traditional cybersecurity practices to encompass the unique challenges of AI/ML systems:



Defense Evasion&	Credential Access&	Discovery&	Collection&	AI Attack Staging	Command and Control&	Exfiltration&	Impact&
8 techniques	3 techniques	8 techniques	4 techniques	4 techniques	1 technique	6 techniques	7 techniques
Evade AI Model	Unsecured Credentials &	Discover AI Model Ontology	AI Artifact Collection	Create Proxy AI Model	Reverse Shell	Exfiltration via AI Inference API	Evade AI Model
LLM Jailbreak	RAG Credential Harvesting	Discover AI Model Family	Data from Information Repositories &	Manipulate AI Model		Exfiltration via Cyber Means	Denial of AI Service
LLM Trusted Output Components Manipulation	Credentials from AI Agent Configuration	Discover AI Artifacts	Data from Local System &	Verify Attack		Extract LLM System Prompt	Spamming AI System with Chaff Data
LLM Prompt Obfuscation		Discover LLM Hallucinations	Data from AI Services	Craft Adversarial Data		LLM Data Leakage	Erode AI Model Integrity
False RAG Entry Injection		Discover AI Model Outputs				LLM Response Rendering	Cost Harvesting
Impersonation &		Discover LLM System Information				Exfiltration via AI Agent Tool Invocation	External Harms
Masquerading &		Cloud Service Discovery &					Erode Dataset Integrity
Corrupt AI Model		Discover AI Agent Configuration					

© The MITRE Corporation

MITRE's ATLAS matrix, similar to its flagship ATT&CK matrix, maps adversary offensive cyber operations as they relate to targeting AI/ML services. Strengthening cybersecurity often begins with understanding how adversaries think and operate.[19]

- **Data poisoning:** Manipulating training data to degrade model performance or to steer outputs towards an attackers' intended outcome.
- **Model theft or tampering:** Gaining unauthorised access to ML models, leading to espionage, theft, or exposing of machine learning-enabled tactics, techniques, and procedures (TTPs).
- **Privacy leakage:** The unintentional disclosure of sensitive or classified information through model output or behaviours.
- **API exploitation:** Attacking the application programming interfaces (APIs) used to interact with models, with the goal of leaking data or disrupting services.
- **Infrastructure attacks:** Targeting underlying compute or storage resources to compromise confidentiality, integrity, or availability.[11]

4.5 Conclusion

AI and machine learning are rapidly expanding in both the scope and complexity of the problems they can help solve. The diversity of technologies and the constantly evolving AI landscape can make it intimidating to begin a project. This is often due to concerns about unfamiliar technical processes that were not traditionally covered in formal education, as well as fears that any solution developed may soon become obsolete. However, structuring a problem with a holistic pipeline, such as the one described in Chapters 3 and 4, can make that first and often most difficult step, getting started, more manageable.

Technically minded NATO professionals are encouraged to familiarise themselves with this pipeline so that they can support AI initiatives as Domain Specialists, even if they have no prior coding experience. It is far more important to develop a solid understanding of what occurs at each stage of the pipeline, and to recognise potential risks such as modelling biases, than it is to be fluent in a specific programming language like Python.

There is a wealth of open-source data available for practice, as well as abundant learning resources such as YouTube, scikit-learn documentation, GitHub repositories, and large language models that can provide live guidance or even generate complete code blocks.

Academia and industry have been building machine learning pipelines for years. It is time for the military to catch up. Therefore, familiarising military personnel with these foundational concepts represents a key step in making NATO an AI-enabled Alliance.

[1] Gen. Devin DL. (2 December 2023). Five methods for data splitting in machine learning. Medium. Retrieved 9 September 2025, from https://medium.com/@tubelwj/five-methods-for-data-splitting-in-machine-learning-27baa50908ed.

[2] Raschka, S. (2014–2023). apriori: Frequent itemsets via the Apriori algorithm. In mlxtend user guide (Version unspecified). Retrieved 9 September 2025, from https://rasbt.github.io/mlxtend/user_guide/frequent_patterns/apriori/.

[3] University of Mannheim. (4 November 2024). Association analysis [Lecture slides]. Data and Web Science Group. University of Mannheim. Retrieved 9 September, 2025, from https://www.uni-mannheim.de/media/Einrichtungen/dws/Files_Teaching/Data_Mining/HWS2024/IE500_DM_07_Association_Analysis.pdf.

[4] Google Developers. (25 August 2025). Classification: Accuracy, recall, precision, and related metrics. Retrieved 9 September, 2025, from https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall.

[5] Ibid.

[6] Ibid.

[7] Fa, A. (23 May 2022). LIME vs. SHAP. Medium. Retrieved 9 September, 2025, from https://medium.com/@afanta/lime-vs-shap-a92623e95c4.

[8] Wang, W. (18 Mar 2020). Bayesian optimization concept explained in layman terms. Medium. Retrieved 9 September, 2025, from https://medium.com/data-science/bayesian-optimization-concept-explained-in-layman-terms-1d2bcdeaf12f.

[9] Gong, L. (12 August 2024). CPU vs GPU vs TPU vs NPU: What are the key differences? Seeed Studio. Retrieved 9 September, 2025, from https://www.seeedstudio.com/blog/2024/08/12/cpu-vs-gpu-vs-tpu-vs-npu/.

[10] Amazon Web Services. (n.d.). What is containerization? Retrieved 9 September, 2025, from https://www.aws.amazon.com/what-is/containerization/.

[11] Stanham, L. (15 May 2025). What is Machine Learning Security Operations (MLSecOps)? CrowdStrike. Retrieved 9 September, 2025, from https://www.crowdstrike.com/en-us/cybersecurity-101/artificial-intelligence/machine-learning-security-operations-mlsecops/.

[12] Bobbitt, Z. (20 September 2021). Validation set vs. test set: What's the difference? Statology. Retrieved September 9, 2025, from https://www.statology.org/validation-set-vs-test-set/.

[13] Carta, S., Consoli, S., Piras, L., & Reforgiato Recupero, D. (February 2021). Structure of the walk-forward approach used for validation [Figure 5]. In Explainable Machine Learning Exploiting News and Domain-Specific Lexicon for Stock Market Forecasting. ResearchGate. Retrieved 9 September, 2025, from https://www.researchgate.net/figure/Structure-of-the-walk-forward-approach-used-for-validation_fig5_349318992.

[14] MathWorks. (n.d.). Choosing the best machine learning classification model and avoiding overfitting. Retrieved 9 September 2025, from https://www.mathworks.com/campaigns/offers/next/choosing-the-best-machine-learning-classification-model-and-avoiding-overfitting.html.

[15] Scrofani, Jim. Classification: Basic Concepts and Techniques. United States Naval Postgraduate School.

[16] Google Developers. (25 August 2025). Clustering algorithms. Retrieved 9 September, 2025, from https://developers.google.com/machine-learning/clustering/clustering-algorithms.

[17] scikit-learn developers. (2023). Choosing the right estimator. In scikit-learn: Machine learning in Python (Version 1.3). Retrieved 9 September, 2025, from https://scikit-learn.org/1.3/tutorial/machine_learning_map/.

[18] Prasad Pandelu, A. (14 December 2024). Day 48: Training neural networks – Hyperparameters, batch size, epochs. Medium. Retrieved 9 September, 2025, from https://medium.com/@bhatadithya54764118/day-48-training-neural-networks-hyperparameters-batch-size-epochs-712c57d9e30c.

[19] MITRE. (n.d.). MITRE ATLAS™ (Adversarial Threat Landscape for Artificial-Intelligence Systems). Retrieved 9 September, 2025, from https://atlas.mitre.org/matrices/ATLAS.

M Cube: © Adobe; Soldier: © Adobe (Generated with AI); Binary Code: © estar 2020 – stock.adobe.com

Chapter 5

Challenges and Limitations of AI in Military Applications

5.1 Introduction

AI-enabled tools are increasingly appearing in planning, operations, and training. They promise faster information processing, improved pattern recognition, and support for complex decisions. However, they also introduce new ways for missions to fail. Models can be trained on incorrect data, produce outputs that appear authoritative but are false, or behave in ways commanders and staff cannot easily explain or challenge. Simultaneously, adversaries can exploit the same technologies to deceive, spread false narratives, or extract sensitive information from seemingly innocuous interactions.

This chapter examines multiple sets of challenges. It addresses the technical limitations affecting whether AI systems can be trained, validated, and safely employed in real-world tasks. It considers how adversaries may deliberately misuse AI against NATO nations within the broader information environment. Finally, it outlines the ethical, legal, and regulatory boundaries that must be respected when integrating AI into military operations. The aim is to provide readers with a realistic understanding of where AI can be beneficial, where it may cause harm, and the conditions necessary before it can be trusted to support operations.

5.2 Technical Challenges

From a technical perspective, the greatest challenges in military AI are not about identifying the most advanced algorithm, but whether the system can be trained, controlled, and understood in real operations. Three technical issues are particularly significant for military organisations. First, assembling suitable quantities of quality data within NATO is often difficult, as the most relevant material may be classified or dispersed across national systems. Second, both the data and the developers' design choices can embed biases that subtly distort operations. Third, AI tools must be sufficiently transparent and explainable for operators to recognise when a model is extrapolating beyond its experience or generating plausible but incorrect content, allowing them to challenge or override its recommendations as necessary. Sections 5.2.1 to 5.2.3 discuss these challenges in detail.

5.2.1 Operational Data Requirements

For military AI, the primary challenge is often not model design but obtaining suitable data. The most relevant training data – operational missions, realistic exercises, intelligence reports, and mission logs – are frequently classified, compartmented, or subject to national caveats. This limits the extent to which data can be pooled, labelled, and shared with developers, particularly in a coalition context. Consequently, many models are trained on peacetime or test-range data that do not fully capture the complexity, friction, and adversary adaptation observed in real operations.[1,2]

Classification further constrains how feedback is gathered and applied. Operators may be unable to export detailed mission data to external vendors or allied nations, hindering model training. Synthetic data can mitigate this, but only if it is carefully designed to reflect realistic conditions, tactics, techniques, and procedures.[3] Commanders should plan on the reality that, absent deliberate curation of representative and releasable datasets at appropriate classification levels, even highly capable models will struggle to translate training performance into operational effectiveness.

5.2.2 Bias as a Technical Risk

Bias in AI systems is a critical issue that can undermine their reliability and operational utility, potentially jeopardising the effectiveness and success of any mission or task supported by the AI.

Bias can enter an AI system at multiple stages, primarily during data collection and training. This is not necessarily malicious; it often reflects the context, timing, and method of original data gathering. Training data that are incomplete, unrepresentative, or focused on particular adversaries or theatres can lead the AI to overestimate risk in those contexts and underestimate it elsewhere. For example, an AI tool trained mainly on state-on-state conflict data may continue to prioritise conventional weapon systems while failing to recognise hybrid activity by non-state actors in the same region. This constitutes *selection bias.*[4]

Commanders and developers should recognise that robust data alone does not preclude bias. Design choices – including feature selection, model structure, and definitions of correct outcomes – can embed distortions during training. For instance, an AI system recommending base-defence postures may place disproportionate weight on historical incident locations. Where these reflect past reporting practices rather than present threat conditions, the system will institutionalise outdated priorities, perpetuating bias rather than correcting it – an effect commonly termed *algorithmic bias.*

5.2.3 Transparency and Explainability

The distinction between transparency and explainability is crucial for building trust in AI-enabled decision-making. Both aim to make systems more understandable but address different needs in practice.[5]

Transparency refers to access to the system's inner workings: code, training data, model architecture and configuration. In principle, a fully transparent AI tool resembles a well-documented machine, where every component can be inspected. For simple models, this may be enough to understand how an output was generated. For modern deep networks, transparency alone is rarely sufficient at the point of decision. A model with billions of parameters may be technically open yet still function as a 'black box' in practice, since neither operators nor legal advisers can feasibly trace how specific inputs led to a particular recommendation.[6]

Explainability, in contrast, focuses on providing a human-understandable rationale for a given output without revealing every internal detail. It connects the model's assessment to meaningful features in the operational context. An explainable system does not merely flag a route as high risk; it demonstrates that this judgment is influenced, for instance, by proximity to recent incidents, deviation from normal traffic patterns, and correlation with previous ambushes.[7] Such explanations allow staff to compare the AI's reasoning with their situational awareness, rules of engagement, and intelligence picture, and to challenge or override recommendations when necessary.

When an AI system lacks both practical transparency and meaningful explainability, an information gap emerges: the system generates prioritisation, indications and warning, or targeting recommendations without revealing the rationale behind them. This gap introduces three distinct risks. First, *alignment:* if the system's underlying logic cannot be examined, it is difficult to confirm that outputs conform to command intent, legal obligations, and political constraints. Second, *trust:* operators may under-rely on a system they do not understand, or over-rely on it if its limitations are opaque – either disregarding valid alerts or accepting flawed outputs uncritically. Third, *accountability:* following an incident, it becomes difficult to determine whether failures stemmed from data quality, model deficiencies, or operator misuse, hindering learning and corrective action.

5.2.4 Hallucinations

Large language models (LLM) introduce an additional complication: *hallucinations*. These systems can generate fluent, confident answers that are factually incorrect, internally inconsistent, or only loosely related to the query. Unlike traditional software, they seldom signal problems through errors. When drafting mission plans, there is a genuine risk that invented details may be incorporated into orders or briefings unless operators treat AI output as a draft requiring verification against trusted sources.

5.3 Adversaries' Intentional Misuse

Even when an AI system is carefully designed and tested, it can be used against friendly forces. For the warfighter, both accidental failure of own AI tools and deliberate adversary misuse are risks, but intentional misuse is increasingly central to opponents' conduct of information warfare and deception.[8]

Generative AI allows adversaries to fabricate highly realistic audio, video and text that mimic trusted authorities, known as *deepfakes*. Deepfaked statements by senior officers, falsified sensor data, and forged imagery can be produced quickly and cheaply. These outputs can be disseminated via social media, messaging apps, and news channels to distort public perception of ongoing operations, civilian harm, or attribution of attacks. Over time, this can erode support for operations and reduce confidence in official strategic communications.[9]

At the tactical level, AI-enabled tools materially increase the effectiveness of vishing and spearphishing attacks. These cyber-enabled techniques use tailored phone calls or emails to initiate intrusions. AI models can generate highly plausible communications using correct terminology, realistic call signs, and current exercise or operational details harvested from open sources. Voice-cloning further enables credible impersonation of commanders or staff officers to request last-minute changes to movements, access, or credentials. The resulting increase in linguistic quality and contextual accuracy makes such attacks harder to detect and easier to scale, amplifying their operational impact.

5.3.1 OPSEC Challenges

Adversaries' intentional misuse also intersects with operations security (OPSEC). An adversary, or a third-party service under their control, could offer an apparently helpful LLM that military personnel use to ask seemingly innocuous questions about capabilities or aspects of mission planning. Each individual query may reveal little in isolation, but when aggregated across a large command, they can expose patterns of deployment, readiness concerns, capability development, or security procedures that become sensitive, or even classified, by aggregation. Research has shown that supposedly anonymised datasets can often be re-identified when combined with auxiliary information, illustrating how difficult it is to guarantee true anonymity once data are collected at scale.[10] In practice, any large collection of prompts and chat logs may become an intelligence source for an adversary if the service is compromised or malicious from the outset.

5.4 Conclusion

AI offers real advantages for planning, operations, and training, but it also introduces new points of failure. Technically, its value depends on access to suitable data, careful management of bias, and the ability of commanders and staff to understand and question what a system is doing. Models trained on narrow or unrepresentative data, or designed without attention to how 'correct' outcomes are defined, will quietly distort priorities and risk assessments. Opaque systems and hallucinating language models can then wrap those distortions in confident prose, making errors harder to spot and harder to trace when something goes wrong.

At the same time, adversaries can exploit AI to undermine decision-making through deepfake-enabled disinformation, highly tailored vishing and spear-phishing campaigns, and the extraction of sensitive patterns from seemingly benign queries. Taken together, these challenges point towards a common requirement: AI must be treated as a powerful but fallible tool. It should be introduced only where there is a clear operational need, fed with carefully curated data, employed with explicit limits and verification procedures, and supported by disciplined OPSEC and user training. Only under those conditions can AI strengthen, rather than weaken, mission effectiveness and trust.

[1] Allen, G., & Chan, T. (2017). Artificial intelligence and national security. Center for a New American Security. Retrieved from: https://www.cnas.org/publications/reports/artificial-intelligence-and-national-security

[2] Ibid.

[3] Michelson, B. (10 January 2024). Data as ammunition: Hyper-personalized warfare in the digital age. Center for European Policy Analysis. Retrieved from: https://cepa.org/article/data-as-ammunition-hyper-personalized-warfare-in-the-digital-age/

[4] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35, https://doi.org/10.1145/3457607.

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

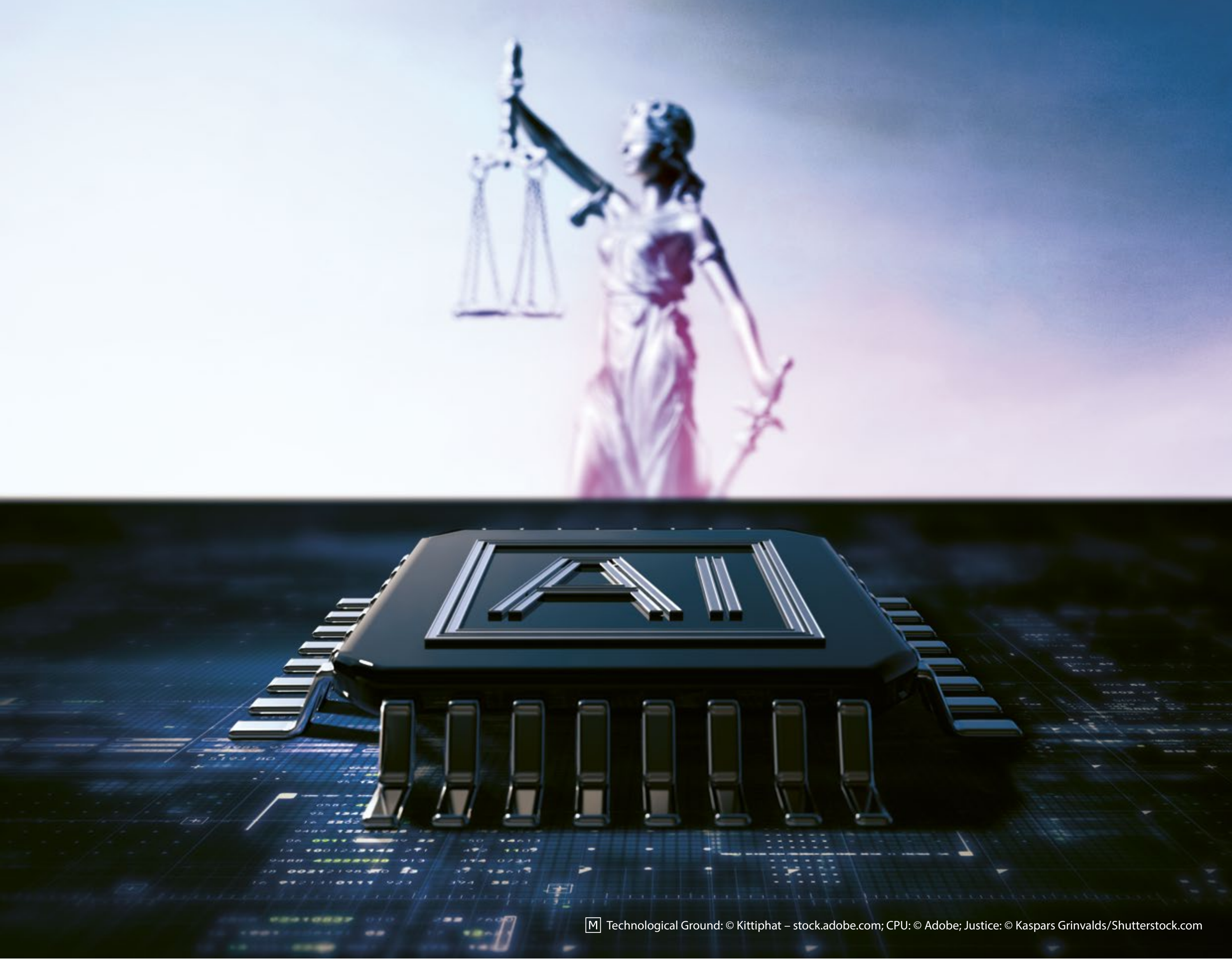
[6] Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence, 267, 1–38, https://doi.org/10.1016/j.artint.2018.07.007.

[7] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 'Why should I trust you?': Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144).

[8] Zhang, W., Sheng, Q. Z., Alhazmi, A., & Li, C. (2020). Generative adversarial networks for cyber security: A survey. IEEE Access, 8, 182673–182695, https://doi.org/10.1109/ACCESS.2020.3037369.

[9] Ferrara, E. (2017). Disinformation and social bot operations in the run up to the 2017 French presidential election. First Monday, 22(8), https://doi.org/10.5210/fm.v22i8.8005.

[10] Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In 2008 IEEE Symposium on Security and Privacy (pp. 111–125). IEEE.



M Technological Ground: © Kittiphat – stock.adobe.com; CPU: © Adobe; Justice: © Kaspars Grinvalds/Shutterstock.com

Chapter 6

Ethical and Regulatory Challenges

6.1 Ethical Challenges

AI compresses decision cycles, at times operating in milliseconds, leaving less room for human judgement. To address these challenges, nations and international bodies are developing frameworks that balance technological innovation with human rights, accountability and safety. NATO organisations, national militaries, and commanders at all levels should ensure adherence to sound ethics and applicable regulations.

6.1.1 Ethical Considerations and OODA Loops

Ethical considerations encompass how AI systems make decisions (autonomy), how responsibility and liability are assigned (accountability), how human rights and humanitarian law are upheld, and how biases are managed. A recurring theme is the OODA loop, the Observe-Orient-Decide-Act cycle developed by US Air Force Colonel John Boyd to describe how fighter pilots can outmanoeuvre opponents by rapidly processing information. Contemporary AI debates often invoke OODA to describe decision cycles in warfare. Researchers caution, however, that while AI can accelerate the OODA loop, focusing solely on speed oversimplifies the model and risks neglecting the critical 'orientation' phase.[1] Orientation involves synthesising information through the lens of experience, culture, and values, something that AI cannot replicate fully. Accelerated decision loops can create pressure to act without adequate reflection, potentially undermining the principles of distinction and proportionality.[2]

6.1.2 Ethical Challenges in AI-Supported Decision Loops

The integration of AI into military decision-making processes introduces complex ethical dilemmas, particularly regarding autonomy, accountability, and compliance with international humanitarian law (IHL). Lethal Autonomous Weapons Systems (LAWS) can detect, select, and engage targets without human control. Critics argue that removing humans from lethal decision-making violates IHL and established moral norms.[3]

As AI systems increasingly influence critical decisions, traditional frameworks for moral responsibility are being challenged.

1. Erosion of Human Responsibility: The shift from human-led to AI-assisted judgment can diminish commanders' ability to fully assume moral responsibility for their decisions, potentially compromising military virtues.[4]

2. Bias Leading to Discrimination: AI systems can perpetuate, or even exacerbate, existing biases, leading to discriminatory practices in targeting or profiling based on race, ethnicity, or other characteristics.[5] The ICRC has warned that the unpredictability and bias of AI systems, particularly in military contexts, can lead to violations of international humanitarian law (IHL).[6]

3. Transparency and Accountability: The opacity of AI-driven models raises significant concerns about trust, accountability, and operational safety, as discussed in section 5.2.3, page 42.[7]

4. Moral Responsibility in Human-AI Teams: Responsibility for civilian harm in human-AI military teams may be contested, increasing the risk of operators becoming detached or suffering moral injury as part of larger human-AI systems.[8]

6.1.3 Human-in-the-loop vs. Human-on-the-loop vs. Human-out-of-the-loop Systems

AI systems can be classified by the degree of human involvement. Human-in-the-loop (HITL) systems require human approval before actions are taken and are appropriate for decisions with legal or ethical consequences. In such systems, operators or commanders directly participate in the system's operation and decision-making process, typically approving, correcting, or providing feedback at critical stages such as training, evaluation, or deployment.

Human on the loop (HOTL) systems operate autonomously but allow human oversight and intervention, with operators able to monitor, intervene, and terminate or override actions within a limited time. In practice, this requires clear authorities delegated to appropriate personnel and realistic training that rehearses such interventions under time pressure.

Humans off the loop, or fully autonomous systems, operate without human input. Military planners often advocate moving off the loop to increase speed, but several experts caution against equating speed with good decisions.[9] AI might handle repetitive tasks, but the orientation phase, evaluating information in context and drawing on experience, requires human judgement. Even in HOTL modes, operators must be trained to recognise when AI suggestions are flawed and they must intervene. Autonomous systems may misidentify targets or apply excessive force. AI trained on biased data might misinterpret behaviour or fail to recognise surrender, for instance. The ICRC warns that AI's unpredictability and bias can lead to violations of international humanitarian law (IHL).[10]

Research on automation bias shows that individuals with moderate knowledge of AI tend to over-trust AI, highlighting the need for training and user interfaces that emphasise uncertainty and require deliberation.[11]

6.2 Dual Nature of AI

AI's dual-use nature permits applications to simultaneously enable both civilian use cases, such as factory automation and self-driving automobiles, and military use cases, such as target identification and battlespace situational awareness. AI systems designed for civilian-specific applications can be transformed or re-engineered to serve military applications, and vice versa. This makes AI a particularly difficult technology to regulate and control. A framework that overlooks this duality risks becoming ineffective. Therefore, any global military AI regulatory framework must take AI's dual nature into account and address these complexities, or risk being compromised.

The Organisation for Economic Co-operation and Development (OECD) AI principles, updated in 2024, call for clear definitions and categories for AI systems and their applications. However, classifying AI technologies as either 'civilian' or 'military' is an oversimplification that fails to capture the complexity of their potential uses.

Instead of simply labelling an AI technology as exclusively civilian or military, it is recommended that definitions be based on a spectrum of potential use cases.[12] For example, an image recognition AI application could be used for factory automation but could also be repurposed for autonomous military targeting. Some experts argue that the classification of AI technology needs to reflect the entire spectrum of use cases. In addition, AI technologies should be classified based on their intended application and the potential risks when repurposed outside their original scope. For example, the same AI technology used for managing supply chains by Amazon carries a low consequence of failure. The worst case may be a delayed shipment. However, when used in a military setting, such as the resupply of troops, the stakes and consequences are much higher. For example, the system may miscalculate a resupply route sending the convoy into an ambush.

6.3 Status of AI Regulation

AI regulation matters because it shapes what capabilities can be acquired, how they must be monitored, and what assurances Allies expect regarding safety, accountability, and data use. The key point is not the legal detail, but that rules are tightening and are not uniform, even amongst NATO nations.

The European Union (EU) and the United States (US) exemplify divergent approaches. The EU is moving towards a comprehensive, risk-based framework for trustworthy AI, while the US is prioritising flexibility and technological competitiveness through a more fragmented, sector-based approach. Military AI is often formally excluded from these civilian regimes, but they still influence industry practices, public expectations, and ultimately, the political constraints under which NATO operations are planned and conducted. The following two pages outline the features and limitations of each approach and includes UN efforts.

6.4 UN Efforts on AI Governance

2024 has seen significant progress by the international community in efforts to achieve multilateral consensus on the use of AI in the military. However, at a time of unprecedented growth in this domain, which poses a risk of uncontrolled proliferation of AI military technologies, much remains to be done. As noted during the UN's First Committee on Disarmament and International Security meeting on 24 October 2023, 'The window of opportunity to enact guardrails against the perils of autonomous weapons and artificial intelligence's military applications is rapidly closing.'[13] It is therefore vital that nations move beyond simply admiring the problem.

On 12 December 2023, the UN General Assembly (UNGA) adopted Resolution 78/241, underscoring the need for the international community to address the ethical and legal concerns associated with the use of Lethal Autonomous Weapons Systems (LAWS).[14]

Until 2023, the UN General Assembly had largely focused on the non-military aspects of AI, calling for international cooperation to ensure that developing states have equitable access to AI technology and its benefits. The military domain was addressed only indirectly through other resolutions. However, 2024 marked a significant development, with the UNGA's

First Committee adopting its first-ever resolution focused specifically on military AI.[15] The resolution reaffirms that existing international legal frameworks, including the UN Charter, International Humanitarian Law, Human Rights Law, and International Criminal Law, are fully applicable to the use of LAWS and related AI-enabled military systems. It highlights growing concerns surrounding the development and proliferation of AI-enabled military applications that could undermine global security and lower the threshold of conflict. A further key aspect of the resolution is its emphasis on responsible and human-centric AI.

The resolution clearly advocates consensus-based international governance of military AI and LAWS and calls for sustained multilateral dialogue and inclusive engagement through further consultations in 2025 under UN auspices.

[M] © olegkruglyak3 – stock.adobe.com

Actor/Instrument	Key Features	Scope & Limitations	Military-AI Implications/Notes
European Union – EU AI Act (Regulation [EU] 2024/1689)	First legally binding framework for AI in the EU. Adopts a risk-based approach that: • Bans practices posing 'unacceptable risk'. • Imposes strict obligations on 'high-risk' systems. • Requires transparency and mandates human oversight for high-risk applications.[16]	Applies within the EU. Explicitly excludes military AI from its scope.	Establishes a strong civilian baseline for trustworthy AI but leaves a regulatory gap for military-specific AI uses.
EU – GPAI Guidelines (2025)	Issued in July 2025, the EU issued guidelines for general purpose AI (GPAI) models, and includes: • A voluntary code of practice. • A template for public training-data summaries.[17]	Non-binding guidelines. Focused on GPAI providers rather than specific sectors.	Influences transparency and accountability expectations, including where GPAI tools are adapted to dual-use or defence contexts.

Actor/Instrument	Key Features	Scope & Limitations	Military-AI Implications/Notes
United States – America's AI Action Plan	The US lacks a single comprehensive AI statute. Since 2023, policy has shifted toward viewing heavy AI regulation as anti-innovation and harmful in the perceived AI arms race with China. America's AI Action Plan emphasises: • Removing 'regulatory barriers' to AI development. • Deregulation and rescinding prior executive actions seen as restrictive.[18]	Fragmented, sectoral approach rather than one overarching AI law. Prioritises innovation and strategic competition.	Fosters rapid development and deployment, including for defence, but raises concerns over gaps in binding safeguards for military AI and LAWS.
United Nations – UNGA Resolutions 78/241 (2023) and 79/62 (2024)	First UN General Assembly resolutions focused on LAWS and military AI. Together they: • Call for guardrails on the military use of AI and autonomous weapons in light of rapid technological growth. • Reaffirm that existing international law (UN Charter, IHL, Human Rights Law, International Criminal Law) fully applies to LAWS and AI-enabled military systems. • Emphasise responsible, human-centric use of AI and the need to address ethical and legal concerns (UNGA Resolutions 78/241 and 79/62).	Global political framework, not a binding treaty regime. Resolutions rely on state implementation and follow-on negotiations; do not set detailed technical standards or verification mechanisms.	Signals growing international concern about military AI and pressure for consensus-based governance. Reinforces expectations that states ensure meaningful human control, legal compliance and risk reduction when developing and employing AI-enabled military systems. Provides a reference point for NATO nations when aligning national policies and Alliance positions on LAWS and military AI.

6.5 Conclusion

The central message of the final two chapters is that commanders should treat AI as a powerful but fallible tool or decision aid, not an oracle. Technical issues such as data quality, bias, hallucinations, and lack of explainability directly determine whether AI tools improve mission outcomes or merely add noise and workload. Ethical and legal considerations, including compliance with IHL and the management of responsibility in human-AI teams, are not abstract considerations; they directly shape how far commanders can safely rely on AI in time-critical situations. Simultaneously, adversaries are exploring how to exploit AI against us through deepfake-enabled information warfare, highly tailored spearphishing, and the aggregation of seemingly harmless queries and metadata into sensitive intelligence.

For NATO and national forces, this implies several concrete imperatives. First, be selective: introduce AI only when a clearly defined decision or process can be improved and where outputs can be verified against trusted sources. Second, design and procure AI with transparency, explainability, and bias management in mind, so that operators can understand, question, and override recommendations rather than work around them. Third, reinforce OPSEC and digital hygiene, assuming that external AI services and the wider information environment are contested spaces vulnerable to adversary exploitation. Used in this way, as a carefully governed and well-understood tool under human command, AI can significantly enhance the Alliance, from the tactical to strategic levels.

Looking beyond individual systems, this handbook as a whole has shown that responsible military AI depends on more than good engineering. It requires a shared understanding of what AI can and cannot do, of how data and models are built and tested, and of the organisational changes needed to integrate them into planning and operations. These are not issues that commanders can solve alone. Governments and the Alliance must provide clear direction on acceptable uses of AI, on the level of human control required for different functions, and on how responsibility is allocated when human-AI teams make mistakes.

Finally, the ethical use and dual-use nature of AI demand greater political and international consensus. Allies will need to converge on minimum standards for safety, testing and human oversight, and to develop common positions on the export, sharing, and employment of dual-use systems that can move rapidly between civilian and military roles. NATO nations should use operational experience to inform national policies and Allied positions in forums such as the EU and UN, pressing for guardrails that are realistic for the battlefield but still protect civilians and uphold existing law. If governments can align on these principles while commanders apply the practical guidance in this handbook, AI can be integrated in a way that strengthens deterrence and defence rather than adding new, poorly understood risks.

[1] Owen. (2021). 'Speeding Up the OODA Loop with AI. A Helpful or Limiting Framework?' JAPCC, https://www.japcc.org/essays/speeding-up-the-ooda-loop-with-ai/.

[2] Amanuel, A. (2024).

[3] Guo, J. (2025). The ethical legitimacy of autonomous Weapons systems: reconfiguring war accountability in the age of artificial Intelligence, https://philpapers.org/rec/GUOTEL#:~:text=It%20argues%20that%20AWS%20fundamentally,a%20reconfiguration%20of%20accountability%20paradigms.

[4] Klaus, M. (2024). Transcending weapon systems: the ethical challenges of AI in military decision support systems. Humanitarian Law and Policy.

[5] Rafia Islam, A. T. (2024). Balancing Power and Ethics: A Framework for Addressing Human Rights Concerns in Military AI.

[6] Laura Bruun, M. B. (2025). 'Constant care' must be taken to address bias in military AI. Humanitarian Law and Policy.

[7] Cheong, B. C. (2024). Transparency and accountability in AI systems: safeguarding wellbeing in the age of algorithmic decision-making.

[8] Devitt, S. K. (2022). Bad, mad, and cooked: Moral responsibility for civilian harms in human-AI military teams.

[9] Ibid, 1.

[10] Ibid, 6.

[11] Lauren Kahn, E. P. (2024). AI Safety and Automation Bias: The Downside of Human-in-the-Loop.

[12] Brundage, M. et al (2018). The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation.

[13] 78th session of The First Committee on Disarmament and International Security. (23 October 2023). Retrieved from: https://press.un.org/en/2023/gadis3725.doc.htm

[14] United Nations General Assembly. (28 December 2023). Resolution 78/241 [Resolution], https://undocs.org/A/RES/78/241.

[15] United Nations General Assembly. (16 October 2024). Resolution 79/62 [Resolution], https://documents.un.org/doc/undoc/ltd/n24/299/16/pdf/n2429916.pdf.

[16] European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council (Artificial Intelligence Act), https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

[17] The official EU policy page on guidelines for providers of general-purpose AI models. (2025). Retrieved from: https://digital-strategy.ec.europa.eu/en/policies/guidelines-gpai-providers

[18] America's AI Action Plan. (2025). Retrieved from Blankrome: https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf

Annex A

Acronym List

AF	Air Force
AGI	Artificial General Intelligence
AI	Artificial Intelligence
ATLAS	Adversarial Threat Landscape for Artificial-Intelligence Systems
AUC	Area Under the Curve
C2	Command and Control
DL	Deep Learning
EDA	Exploratory Data Analysis
EU	European Union
F1	F1 Score
FN	False Negative
FP	False Positive
GPS	Global Positioning System
HITL	Human-In-The-Loop
HOOTL	Human-Out-Of-The-Loop
HOTL	Human-On-The-Loop
ICRC	International Committee of the Red Cross
IHL	International Humanitarian Law
ISR	Intelligence, Surveillance, and Reconnaissance
IT	Information Technology
LAWS	Lethal Autonomous Weapons Systems
LIME	Local Interpretable Model-Agnostic Explanations
LLM	Large Language Model
MDO	Multi-Domain Operations
ML	Machine Learning
MLSecOps	Machine Learning Security Operations
MSE	Mean Squared Error
MUM-T	Manned-Unmanned Teaming
NATO	North Atlantic Treaty Organization
NLP	Natural Language Processing
OECD	Organisation for Economic Co-operation and Development
OODA	Observe-Orient-Decide-Act
OPSEC	Operations Security
PCA	Principal Component Analysis
R2	Coefficient of Determination (R-squared)
RFE	Recursive Feature Elimination
RMSE	Root Mean Squared Error
ROC	Receiver Operating Characteristic
SAM	Surface-to-Air Missile
SHAP	Shapley Additive Explanations

SMOTE	Synthetic Minority Over-sampling Technique
TEVV	Test, Evaluation, Verification, Validation
TN	True Negative
TP	True Positive
UAS	Unmanned Aerial System
UN	United Nations
UNGA	United Nations General Assembly

Annex B

About the Authors

Colonel (ret.) Antonios Chochtoulas

Hellenic Air Force, Cyber Subject Matter Expert, JAPCC (former)

Colonel Chochtoulas graduated from the Hellenic Air Force (HAF) Academy in 1999, with a degree in Logistics.

He holds a Master of Science in Computer Science from the Hellenic Open University, and his subject matter expertise is in the Areas of Information Systems Security and Database Design and Administration. He initially served as a programmer and thereafter as a Database and System Administrator of HAF's proprietary Logistics Information System in Elefsis Air Base. His previous assignment was at the HAF Supply Depot as Director of the IT Department. While at the JAPCC, he was the Cyberspace SME.

Now he works as an Information Security Officer at the European Union Agency.

Lieutenant Colonel Nakul Nayyar

Royal Canadian Air Force, Cyber Subject Matter Expert, JAPCC

Lieutenant Colonel Nakul Nayyar joined the Royal Canadian Air Force in 2004 as a communications and electronics engineer.

He most recently served as the Canadian Communications & Electronics Defense Attaché in Washington, D.C. (2020–2024). In this role, he provided expertise, assistance, and advice in the information and Cyber Defence domains affecting the US-Canada bilateral defence sector partnership. Concurrently, he served as Permanent Secretary of the Combined Communications Electronics Board (2021–2024), leading modernisation efforts for Five Eyes (FVEY) digital interoperability between Australia, Canada, New Zealand, the United Kingdom and the USA.

Academically, Lieutenant Colonel Nayyar holds a bachelor's degree in electrical engineering from the University of Toronto and two master's degrees in business administration and defence studies from the Royal Military College of Canada. He currently serves as a Cyber SME at JAPCC, leveraging his experience to enhance NATO's cyber capabilities.

Major Lucas J. Stensberg

US Space Force, Space and Cyber Subject Matter Expert, JAPCC

Major Stensberg is a space and cyber SME in the JAPCC's C5ISR & Space branch, furthering the Alliance's understanding of the two domains via concept development, exercises, wargames, doctrine, and training. Before this role, he served in the US Space Force's Enterprise Talent Management Office, and prior to that, as a Cyber Operations Planner at Headquarters 16th Air Force. There, he aligned strategies with US Cyber Command and notably the newly stood-up US Space Command. Other previous assignments include Flight Commander of Tactical Communications for the 485th Intelligence Squadron, managing C4ISR capabilities for 29 partner nations and over 900 intelligence analysts, as well as Integrated Project Management supporting the 694th ISR Group in Osan, Republic of Korea. Major Stensberg commissioned as a Cyberspace Operations Officer in 2016 from the United States Air Force Academy. His formative years were spent undergoing cyberspace warfare training at Keesler Air Force Base.

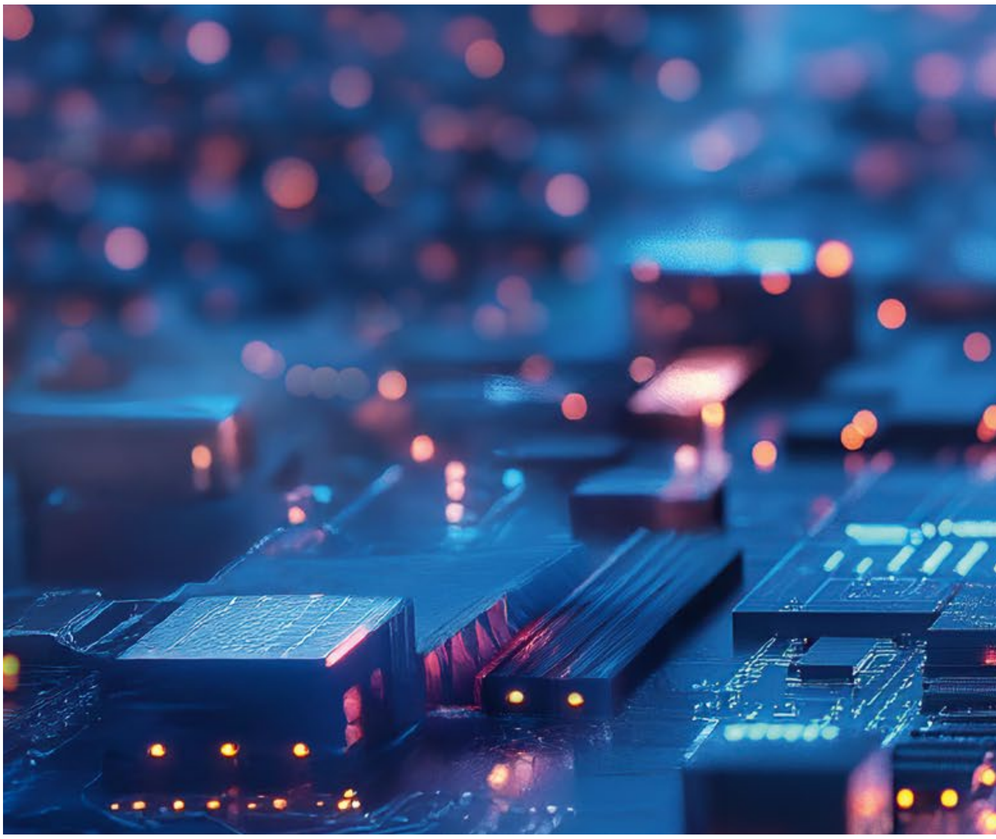
Laura Samsó Pericón

Researcher and Strategist, Airbus

Laura Samsó Pericón is a researcher and strategist focused on cyber resilience, AI-enabled autonomy, and trusted human-machine teaming, including MUM-T and high-altitude platform concepts.

With 15+ years of international experience across civil and defence environments and multi-domain operational settings, her work examines AI-enabled decision-making in manned-unmanned teaming (MUM-T) contexts and cognitive and neuro-adaptive technologies. She previously served as an Executive Vice President and co-founded a technology venture, adding business and leadership depth to her technical expertise. She has also contributed to regulatory and standards bodies, alongside expert groups and international defence forums addressing autonomy, cyber resilience, and emerging technologies.

As Founder of Synarea Insights, she advances an assurance-by-design approach – integrating cyber, autonomy, AI, HMI/BCI, ethics, and secure human integration – to strengthen trust and mission performance in future capabilities in defence operations. Her background in Electronics and Aerospace, plus expedition leadership in high-challenge field environments, informs her perspective on resilience, trust and decision-making under uncertainty.



Joint Air Power Competence Centre

von-Seydlitz-Kaserne
Römerstraße 140 | 47546 Kalkar (Germany) | www.japcc.org