

IMPERIAL



SOVEREIGN AI IN DEFENCE

**Strategic Autonomy and
Collaborative Opportunity**

Adele Jashari¹, David Shrier^{*1,2} & Aldo Faisal^{*1,3}

(1) Imperial College London, UK (2) Massachusetts Institute of Technology, USA (3) Universität Bayreuth, Germany

*corresponding authors david.shrier@imperial.ac.uk & aldo.faisal@imperial.ac.uk

Table of Contents

Executive Summary.....	04
1. Introduction.....	08
2. Defining Sovereign AI in the Defence Domain.....	11
2.1 Strategic Dimensions of Sovereign AI.....	11
2.2 Strategic Imperative.....	13
3. Strategic Drivers.....	14
3.1 Geopolitical instability and Grey-Zone Conflict.....	14
3.2 Key Drivers of change.....	14
3.3 Conflict Acceleration and Diffusion.....	15
3.4 Implications for capability development.....	16
3.5 Illustrative Examples from the Russia-Ukraine Conflict and Emerging Technologies.....	19
3.6 Non-Compromising Interoperability.....	20
3.7 Expansion of Defence Concept: Blended Warfare Across Domains.....	21
4. Rationale for Sovereign AI.....	23
4.1 Operational Assurance and Mission Integrity.....	24
4.2 Legal and Ethical Responsibility.....	25
4.3 Strategic Autonomy and Freedom of Action.....	26
4.4 Alliance Interoperability on Sovereign Terms.....	26
4.5 Domestic Industrial and Economic Security.....	29
5. Operational Domains Requiring Sovereign AI.....	31
5.1 Intelligence, Surveillance and Reconnaissance (ISR).....	31
5.2 Command Decision Support.....	33
5.3 Cyber Operations - Defensive and Offensive.....	34
5.4 Tactical Autonomy and Embedded Inference.....	35
5.5 Cognitive Security and Information Operations.....	36
6. Economic Modelling and Feasibility.....	38
6.1 Infrastructure Requirement for Sovereign AI.....	38
6.2 Emerging Sovereignty Opportunities in Neuromorphic and Hybrid Compute.....	41
6.3 Model Sovereignty.....	42
6.4 Secure-by-Design Processors and Hardware Assurance.....	43
6.5 Beyond the capital cost.....	45
6.6 Constraints and dependencies.....	48
6.7 Strategic Enablers of Sovereign AI - Talent, Clearance, and Workforce Development.....	49
6.8 Data Advantage in Defence: From Dark Data to Sweet Spot Models.....	51
6.9 Strategic Integration, UK MOD: Insights from the Sovereign AI Initiative.....	52
6.10 Safety, Theory of control and national Assurability.....	54

7. Strategic Synthesis and Posture Recommendation..... 57
7.1 Operationalising Sovereignty: The Role of Metrics.....60
7.2 Core Dimensions of Sovereign AI Metrics.....62
7.3 Implementation Role of the Directorate.....63

8. Refined Hypothesis for Mission-Driven, Modular Sovereignty..... 64
8.1 Rationale for the Hypothesis - UK MOD.....64
8.2 Evaluation and Testing Pathways.....65
8.3 Strategic Deterrence, Legal Flexibility, and Adversarial Asymmetry.....67
8.4 Forward Looking Utility.....68

9. Risk Landscape and Adversarial Dependencies..... 69

10. Comparative International Postures..... 71

11. International Collaboration and Coalition Sovereignty..... 73
11.1 Multilateral Structures as Vehicles for AI Norm-Setting.....73
11.2 Priority Areas for Technical and Doctrinal Collaboration.....74
11.3 Diplomatic and Strategic Considerations.....76
11.4 From Coalitions of Convenience to Coalitions of Sovereignty.....77
11.5 Leveraging Partnerships and Multilateral Collaboration.....78
11.6 Export Controls and Technological Assurance in Sovereign AI.....80

12. Conclusions.....81
12.1 Implications for Policy and Capability Development.....81
12.2 Strategic Imperatives and the Cost of Inaction.....81

About the Authors..... 83

Glossary of Terms and Abbreviations..... 84

Executive Summary

We present a strategic framework for achieving mission-driven and modular AI sovereignty, a model designed to offer practical guidance for governments, defence institutions, and policymakers seeking to preserve legal authority, operational autonomy, and strategic freedom of action as artificial intelligence becomes increasingly embedded in critical national security infrastructure. This framework builds on the principles developed through Imperial's [Trusted AI Alliance](#) and extends our earlier work on [Sovereign AI and National AI Policy \(2025\)](#) into the specific demands of the defence and security domain.

While the United Kingdom and its Ministry of Defence (MOD) are used as illustrative case studies, the frameworks and posture model outlined in this paper are globally relevant, applicable to any state facing the accelerating convergence of digital automation, strategic decision-making, and sovereign command responsibility.

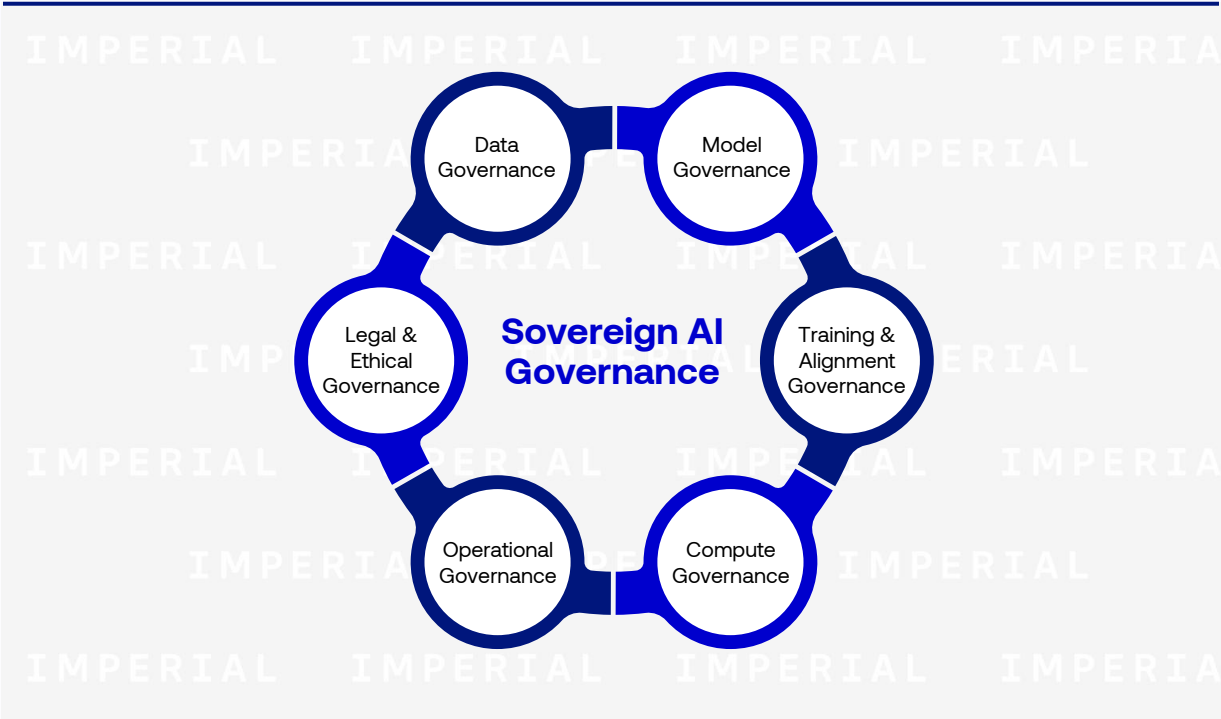
At the heart of this framework is the recognition that not all AI systems demand the same degree of sovereign control. A mission-driven, modular approach to Sovereign AI focuses attention and resources on those capabilities where the consequences of external interference, misalignment, or legal ambiguity could directly undermine national security, lawful command, or strategic autonomy. This includes domains such as targeting, cyber defence, command decision support, and autonomous battlefield operations. Across all applications, the principle remains the same: AI systems must be subject to governance, assurance, and oversight structures that ensure they operate in alignment with national objectives, legal standards, and strategic intent.

This approach recognises that total technological self-sufficiency is neither realistic nor necessary in today's interconnected global environment. Instead, it calls for trusted assurance over mission-critical systems, ensuring that even when global supply chains or allied technologies are involved, the state retains the ability to govern, validate, and, when required, intervene to uphold national authority.

In this context, Sovereign AI refers to the assured capacity of a state to maintain control over the AI systems on which its national security increasingly depends. It is not defined by complete technological self-sufficiency, but by the ability to govern, validate, and, when necessary, intervene in the operation of AI systems, preserving the authority to act and decide under national command, even amidst complex global supply chains, alliance frameworks, and contested information environments.

AI systems must not only deliver performance, but remain governable, auditable, and adaptable, capable of functioning predictably in contested environments, under degraded communications, and amidst shifting legal or operational conditions.

To achieve this, sovereignty must be delivered through a coherent governance framework spanning six core interdependent dimensions. Each plays a distinct role in safeguarding the AI systems that underpin national defence, ensuring they remain under assured national control even in the face of technological, environmental, or geopolitical disruption.



Crucially, Sovereign AI is sustained not by technology alone, but by the presence of institutional capacity and trusted human oversight across these domains. The ability to govern, re-align, and safely control AI systems in real-time, particularly in high-consequence or degraded environments, requires a cleared, strategically aligned workforce embedded within defence and security institutions. Without this human capability, technological assets cannot deliver meaningful sovereignty: AI without sovereign stewardship is not an advantage, but a vulnerability.

We identify a near-term opportunity for strategic advantage: the development of high-fidelity, domain-specific Sovereign AI “sweet spot” models based on sovereign defence datasets such as sonar, Radio Frequency telemetry and mission logs. These models offer operational alignment and legal control beyond what is achievable with general-purpose commercial systems. For countries with access to these national data sources, this is a defensible frontier, provided there is investment in trusted compute, legal assurance, and curated model development pipelines.

While this paper applies universally, the UK case illustrates specific structural challenges:

 A limited pipeline of cleared AI professionals	 Fragmented compute infrastructure	 Over-reliance on commercial assurance
---	--	---

Addressing these will require coordinated action across the Digital Defense ecosystem (eg. MOD Strategic Command, DSTL,

Defence Digital, the Defence AI Centre) and cross-government leadership.

Sovereign AI demands developmental agility:

the capacity to reclassify, retrain, and reassert authority over AI systems as threat environments evolve. This agility must be embedded into defence doctrine, exercised through wargaming, and resourced as a standing institutional capability. Without the ability to continuously govern and adapt AI systems under operational pressure, sovereignty risks becoming brittle, eroded by technological inertia, strategic surprise, or adversarial manipulation. It is this need for adaptive and assured control that makes sovereign AI far more than a matter of defending technological advantage.



It is fundamentally about preserving the authority to act, decide, and lead in moments of geopolitical consequence. In an era where strategic power is increasingly exercised through code, the ability to govern that code will shape the boundaries of national freedom.

Current debates on autonomous weapon systems have rightly focused attention on the legal and ethical challenges of AI in the

use of force. Yet these systems represent only the most visible tip of a much broader transformation. Long before automation reaches the point of weapon release, AI already shapes what threats are perceived, what choices are surfaced, and how options are framed for human decision-makers. Without sovereign control over these upstream processes, states risk losing the ability to fully explain or justify the decisions taken in their name.

To deliver credible Sovereign AI, governments must act across these key strategic priorities.

Secure sovereign compute capacity by scaling trusted, resilient infrastructure both at national and tactical levels

Build a security cleared, skilled AI workforce through dedicated career pathways, public-private partnerships, and security clearance reform

Embed rigorous assurance, testing, and legal oversight into AI development, particularly for high-consequence systems

Strengthen national industrial capacity to protect sensitive AI assets, supply chains, and intellectual property

Shape AI governance within alliances, ensuring interoperability without surrendering sovereign decision authority

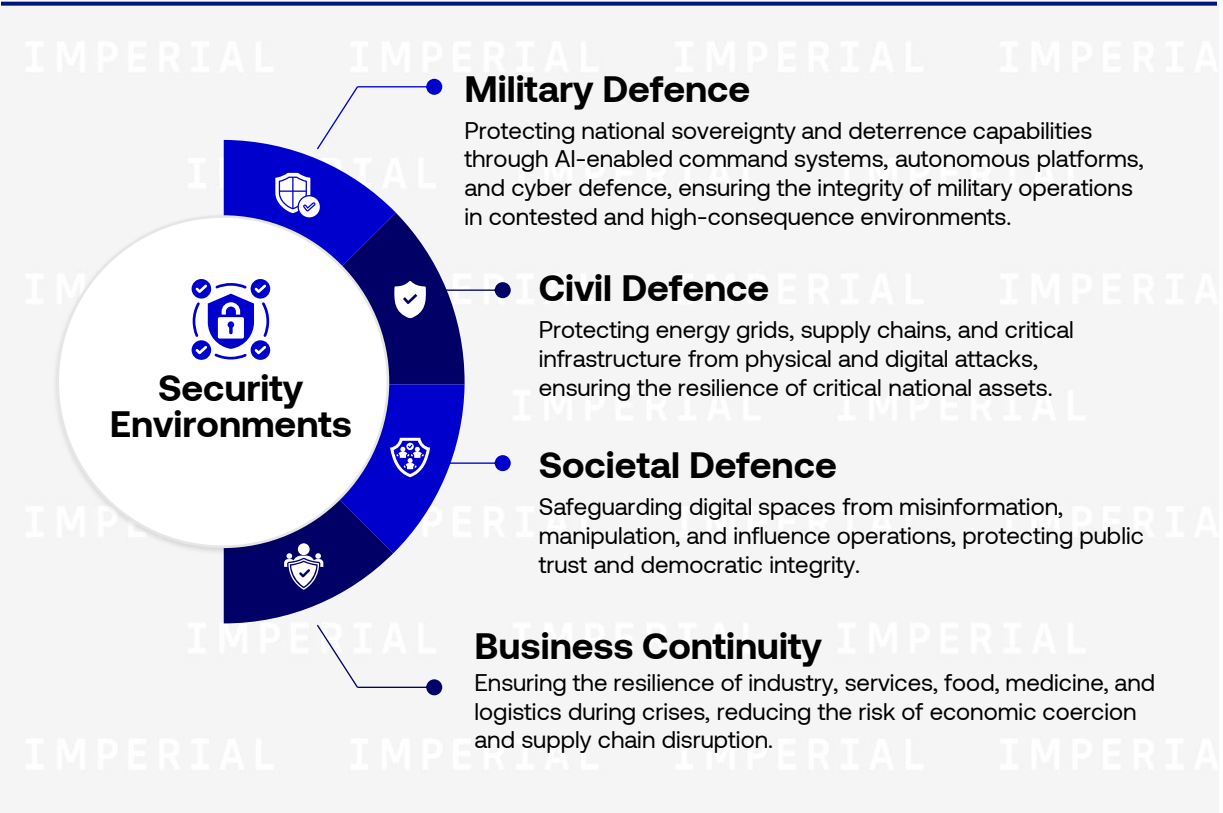
Thought leadership in cooperation across Sovereign AIs is a soft power

Alongside these priorities, the paper introduces the concept of uncompromised interoperability, the principle that AI-enabled systems must be able to operate within allied and coalition structures without surrendering sovereign control over decision-making authority, legal accountability, or system governance. As AI becomes increasingly embedded in joint operations, preserving this ability to cooperate without compromising national authority will be essential to sustaining both alliance credibility and operational independence.

In addition, Sovereign AI must be resilient across blended security environments. As

the boundaries between military, civil, social and business continuity domains continue to converge across emergency response, infrastructure protection, and mobilisation planning, the principles of sovereign control must extend beyond traditional warfighting systems to encompass this broader security landscape.

Sovereign AI will not emerge by default. It requires deliberate investment, institutional reform, and integrated leadership across defence, industry, and government to ensure that nations retain the authority to act, decide, and lead in an era shaped by accelerating automation.



1. Introduction

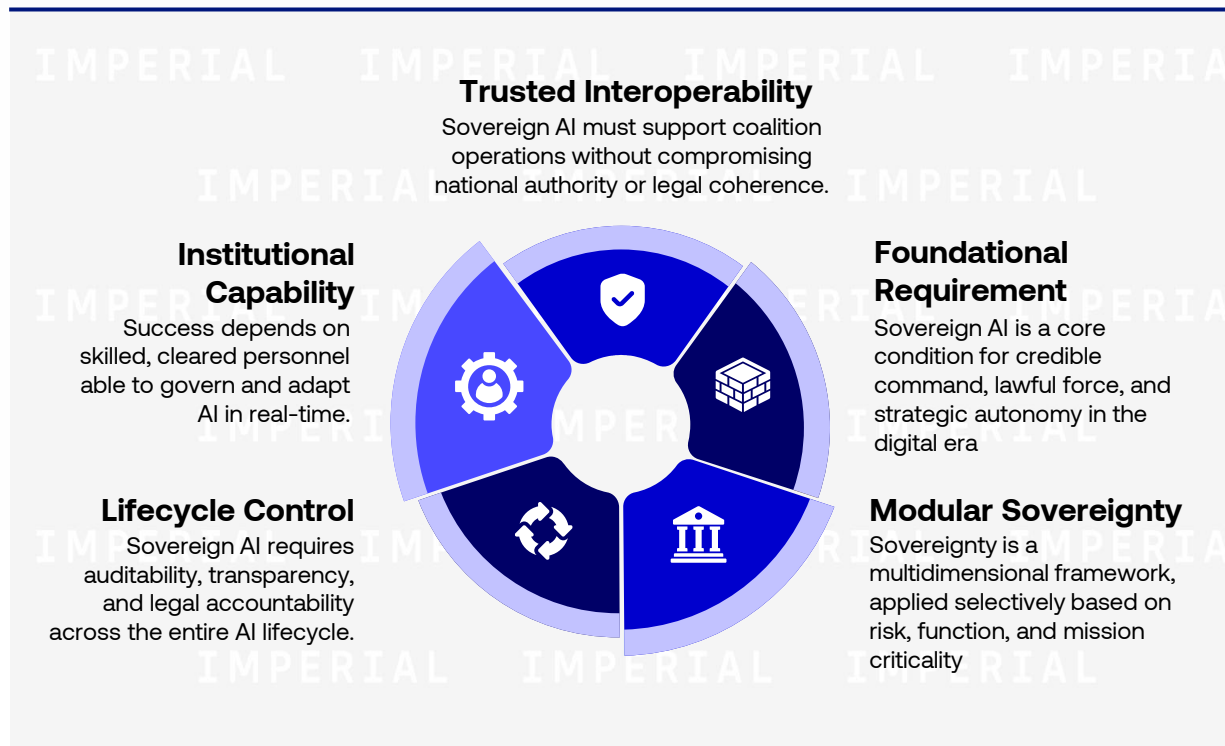
Artificial intelligence is reshaping the global landscape, including the way defence and security institutions perceive threats, support decision-making, and manage complex operations. While its adoption remains uneven, AI is increasingly being integrated into systems that assist with analysis, planning, coordination, and, in some cases, autonomous action under operational constraints.

As the AI transformation accelerates, a strategic question confronts all states: not whether to adopt AI in defence as we already passed that point, but how to do so in a way that preserves national authority, legal accountability, and operational credibility? This is the central concern of this white paper.

As outlined in our previous publication, sovereignty in AI is not a binary status but a strategic spectrum, one in which states must balance the imperative for self-reliance with the practical need for global AI interoperability and collaboration. **The term**

“Sovereign AI” refers to a nation’s ability to develop, control, and regulate artificial intelligence systems independently, ensuring that these technologies align with its national security, economic interests, and ethical values.

In this defence-focused paper, we extend that foundation to **define sovereign AI as the assured capacity to govern, deploy, and, where necessary, override the AI systems used to inform or execute defence functions.** This includes ensuring that critical AI capabilities remain aligned with national objectives, auditable under domestic law, and governable in real time, particularly in moments of military escalation or geopolitical crisis. Sovereign AI, in this context, is not about autarky, but about preserving the authority to act, decide, and lead, even in the face of technological dependency, alliance complexity, and contested information environments. Key running themes in this paper include:



This paper argues that **Sovereign AI is a foundational requirement for credible national defence in the digital era.** We set out the operational domains in which sovereignty is essential, the criteria by which it should be calibrated, and the institutional structures required to enforce it. We use the United Kingdom's evolving approach as an illustrative case study, however our framework applies broadly to any open society state seeking to maintain operational autonomy while engaging in trusted international collaboration. As the adoption of AI accelerates, the window to shape its governance narrows. **Sovereignty must therefore be treated not as a policy aspiration, but as an urgent operational requirement.**

As our paper's exemplar country, it is clear that the United Kingdom, as a globally committed middle power and nuclear-armed state with deep alliance structures, cannot afford to approach this transformation passively. Within the Ministry of Defence (MOD) and comparable institutions globally, AI systems are increasingly integrated across the full spectrum of operational, strategic, and institutional activities. These

technologies are shaping how forces are trained, how missions are planned, how logistics are coordinated, and how operational decisions are supported.

Consistent with the United Kingdom's legal, ethical, and doctrinal principles, there remains a clear commitment to human-centred AI systems designed to enhance professional judgement, support lawful command authority, and preserve institutional accountability. This approach avoids the pursuit of fully autonomous or agentic warfare models in favour of architectures that reinforce meaningful oversight and strategic control.

The Trusted AI Alliance presents this white paper to provide structured guidance for governments seeking to navigate the demands of AI sovereignty across defence, civil, and societal domains. Using the United Kingdom's evolving experience as an illustrative case study, we examine practical pathways to achieving Sovereign AI capabilities that uphold ethical standards, strategic autonomy, and trusted international collaboration.

We specifically address:



The defining characteristics of Sovereign AI across defence, civil resilience, and strategic infrastructure, grounded in the six dimensions of governance.



How sovereignty should be calibrated based on mission criticality, legal obligations, operational risk, and strategic consequence.



The operational domains where Sovereign AI is essential, and the risks associated with non-sovereign control in high-consequence applications.



The industrial, institutional, and infrastructural foundations required to build and sustain Sovereign AI capabilities over time.



The emerging challenge of hardware-software co-design and the strategic significance of sovereign compute, trusted silicon, and secure-by-design architectures.



The necessity of embedding AI safety, interpretability, and assurance as core sovereign functions, not technical afterthoughts.



The metrics, thresholds, and classification tools required to assess, verify, and enforce sovereign control over AI systems.



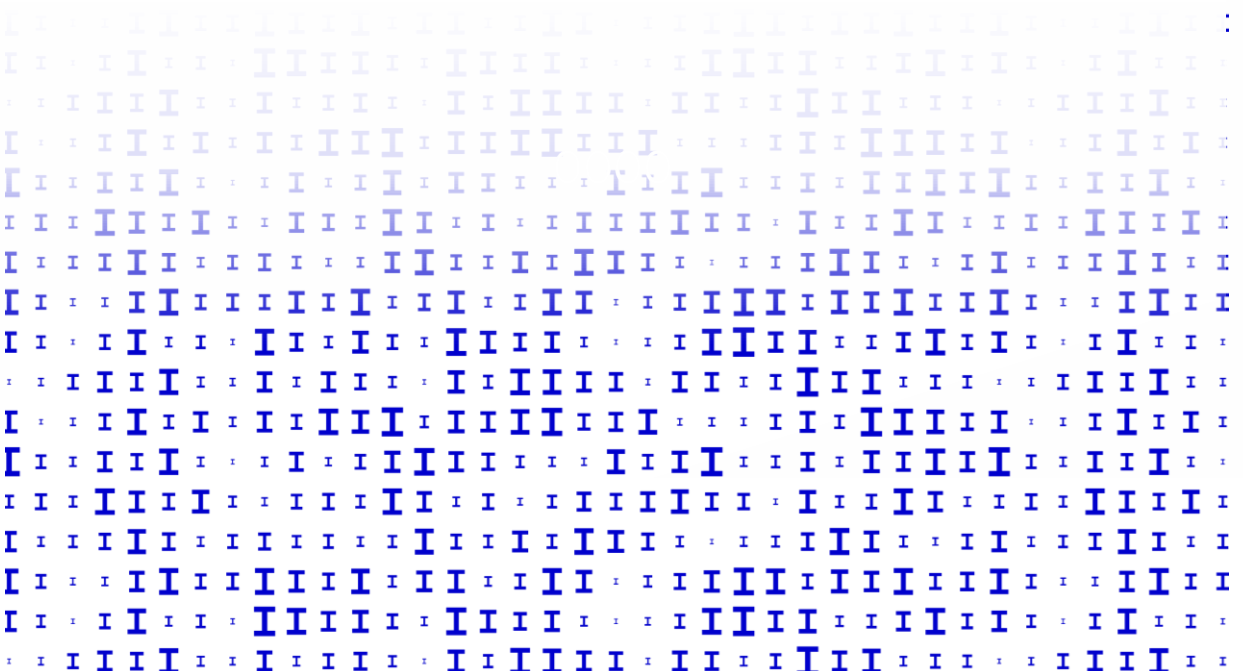
How to enable trusted interoperability with allies through modular integration and shared assurance frameworks, without compromising national control or legal accountability.



The strategic consequences of inaction, and the structural commitments required to preserve decision authority, legal legitimacy, and command credibility in an era of automated conflict.

Our analysis is grounded in publicly available government publications, strategic doctrine, international law, national AI and digital strategies, including comparative insights from peer nations. It is structured to support

policymakers, planners, and operational leaders in building Sovereign AI ecosystems that are resilient, ethically grounded, and strategically credible in an era defined by accelerated technological evolution.



2. Defining Sovereign AI in the Defence Domain

The term “Sovereign Artificial Intelligence” has become increasingly prominent in national security, defence, and digital policy discourse. While the concept is introduced earlier in this paper, its precise meaning warrants further clarification, particularly where national resilience, legal authority, and operational credibility are concerned. In the context of defence, Sovereign AI refers to the assured capacity of a state to govern, deploy, and, where necessary, override the AI systems critical to national security, ensuring they remain under domestic legal control and strategic alignment.

Sovereignty in the Defense and security context does not necessarily demand that every component, from silicon to software, from all human talent to all data, be domestic (much less government manufactured). Rather, it requires the capacity to assure, validate, and govern the AI systems relied upon for mission-critical functions. This includes the ability to verify the design and objectives of AI models, ensuring that they align with national security requirements, ethical standards, and operational priorities. However, this need not imply direct control over all training data, provided that robust ‘black box assurance’ methods are in place.

2.1 Strategic Dimensions of Sovereign AI

Achieving a balanced and effective Sovereign AI posture necessitates a comprehensive framework built upon six interdependent dimensions: Data Governance, Model Governance, Training and Alignment Governance, Compute Governance, Operational Governance, and Legal and Ethical Governance. Each pillar addresses specific facets of AI development and deployment, ensuring that AI systems are robust, secure, and aligned with national interests.

Collectively, these six dimensions form a comprehensive framework that supports the development and deployment of Sovereign AI systems ensuring that AI technologies serve national interests, uphold democratic values, and maintain public trust. However, sovereignty should not be seen as a high control, closed loop requirement in every context. Rather, it should focus on ensuring that national defence organisations can use their AI-enabled capabilities without the risk of external control or interference, even

when those systems incorporate externally sourced components. This requires a risk-based, layered approach that prioritises sovereignty where it matters most, while allowing for flexibility and innovation in less critical areas.

In practical terms, sovereignty begins with control over data, which forms the foundation for all downstream assurance. Without confidence in the provenance, curation, and security of data, no model or inference process can be fully trusted or audited. This control must extend across the governance pipeline. Among these, alignment governance, the ability to shape how AI systems internalise goals and constraints, will become increasingly strategic as AI-driven alignment of AI itself evolves. Sovereign AI is not only about who builds the system, but who retains the capacity to govern its behaviour, adapt it in real time, and ensure its outputs remain under lawful national authority.

01. Data Sovereignty

This dimension concerns the ability to govern the origin, access, security, and legal classification of the data used to train and operate AI systems. It focuses on ensuring that sensitive defence datasets such as operational telemetry, ISR feeds, and classified mission data are curated, protected, and remain under national control. While data quality influences outcomes, issues of bias and behaviour are addressed at other layers of governance.

02. Model Sovereignty

Model sovereignty refers to control over the architecture, weights, parameters, and technical design of AI models. It ensures that national authorities retain the ability to inspect, modify, and understand how models process inputs and generate outputs, particularly in high-consequence applications. This dimension underpins explainability, technical robustness, and the ability to align or constrain system behaviours as required.

03. Training and Alignment Sovereignty

This dimension governs the process by which AI systems are trained, fine-tuned, and aligned with strategic, legal, and ethical objectives. It includes the design of reward functions, safety tuning, the use of human or AI feedback loops e.g. Reinforcement Learning from Human Feedback (RLHF), Reinforcement Learning from AI Feedback (RLAIF), and alignment with mission-specific values. It is the layer at which the control problem is addressed, ensuring AI behaviours reflect national policy and remain adaptable over time.

04. Compute Sovereignty

Compute sovereignty ensures national control over the hardware, cloud infrastructure, supply chains, and energy dependencies used to train, deploy, and operate AI models. It encompasses both centralised data centre capability and the need for deployable, low-power edge AI suitable for contested, communications-denied environments. Without this layer, AI systems remain vulnerable to external dependency or denial.

05. Operational Governance Sovereignty

Operational governance focuses on how AI systems are deployed, monitored, and governed during live operations. It includes mechanisms for human oversight, intervention, and rollback; the embedding of failsafe behaviours; and the ability to maintain lawful and ethical operation under degraded conditions. This dimension ensures that AI systems remain accountable and governable when deployed in dynamic or contested environments.

06. Legal and Ethical Sovereignty

The final dimension ensures that AI-enabled decisions comply with national law, international humanitarian law (IHL), and ethical principles. It safeguards legitimacy, supports alliance trust, and preserves the moral and legal standing of national defence actions. Legal and ethical sovereignty must be embedded not only in policy but in the technical and operational layers of AI system design and deployment.

Cross-Cutting Principle: Assurance

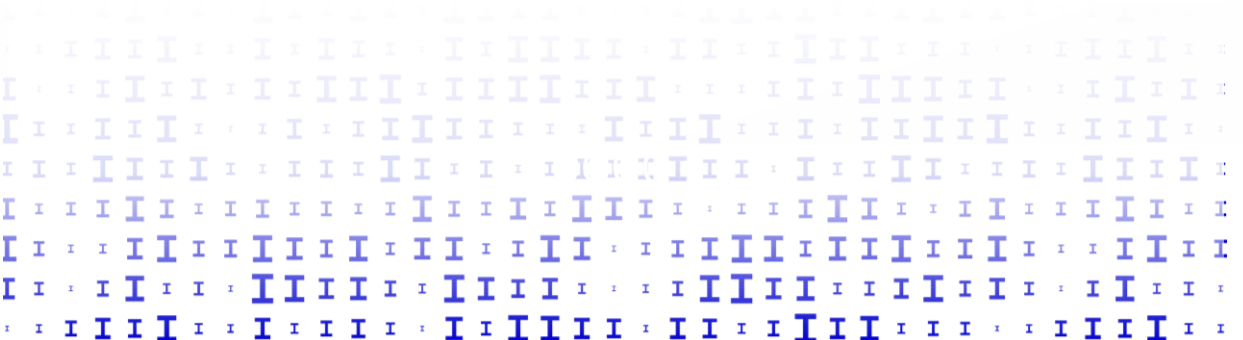
While not framed as a standalone dimension, assurance, encompassing safety verification, bias detection, adversarial robustness, and auditability, must be embedded across all six domains. Sovereign AI is not credible without the continuous capacity to test, validate, and, where necessary, constrain the systems deployed.

2.2 Strategic Imperative

In modern operations, the resilience and sovereignty of front-line AI capabilities depend not only on the performance of weapons systems or ISR platforms, but equally on the logistics, data networks, decision-support tools, and administrative infrastructures that sustain them. Sovereign AI must therefore encompass this entire defence technology ecosystem, recognising that vulnerabilities, dependencies, and decision-framing risks can emerge from any layer, not solely from combat applications.

This broader approach to sovereignty reflects the reality that AI systems are not isolated technical components but integral parts of complex, interconnected defence ecosystems. As such, the ability to assure,

validate, and adapt these systems is a strategic imperative, ensuring that the decisions and operations vital to national security remain under lawful, ethical, and nationally accountable control. As the [UK MOD's Defence Artificial Intelligence Strategy \(2022\)](#) makes clear, this approach must be firmly grounded in responsible innovation, legal compliance, and alliance interoperability, reflecting the broader imperative that democratic nations must lead by example, ensuring that AI capabilities enhance human authority, reinforce democratic values, and preserve the strategic freedom to act in contested operational, informational, and geopolitical environments.



3. Strategic Drivers

The case for Sovereign AI in defence emerges from a confluence of global, national, and institutional pressures that make technological control both a strategic necessity and a policy urgency. These

drivers are visible across key areas such as the international threat environment, technological geopolitics, alliance dependence, and digital vulnerability within defence infrastructure.

3.1 Geopolitical Instability and Grey-Zone Conflict

The nature of warfare is undergoing a profound transformation. Traditional divisions between war and peace have become increasingly blurred, with modern conflicts now characterised by continuous competition across multiple domains. Grey zone operations, hybrid threats, and information warfare are now as integral to conflict as conventional military operations. Adversaries seek to achieve strategic effects below the threshold of open warfare, exploiting political, economic, cyber, and information levers to undermine and destabilise opponents without necessarily engaging in full-scale combat.

A defining feature of this new environment is the speed at which events unfold. The actor that can outpace its opponent, whether in decision-making, deployment, or adaptation, is more likely to gain quick wins and eventually succeed. Control of information, both in securing one's own data and manipulating the adversary's perception, is becoming as critical as physical control of

territory. Moreover, non-kinetic capabilities such as cyber attacks, economic coercion, and influence operations are now capable of achieving strategic objectives that previously would have required military force.

The Russia-Ukraine war has provided a stark illustration of these dynamics. Although conventional forces, such as artillery, armor, and infantry, have remained decisive on the battlefield, the conflict has also demonstrated the centrality of information warfare, resilient command-and-control networks, and flexible, adaptive tactics. Russia's early failures to integrate these dimensions into its operations and Ukraine's innovative use of Western-supplied technology, decentralised decision-making, and agile responses have underscored that future success demands forces that are not only physically capable but also digitally superior, agile, and resilient across all domains. AI drone technology, in particular, has seen a number of notable advances in this conflict.

3.2 Key Drivers of Change

Several interrelated forces are reshaping the future conflict environment. The first is the accelerating pace of technological change, particularly in the fields of artificial intelligence, autonomy, and hypersonic weapons. These technologies are fundamentally altering the speed, scale, and

character of modern warfare. Capabilities that once took decades to mature are now emerging within years or even months, compressing adaptation cycles and shifting advantage decisively toward those actors able to integrate, iterate, and deploy at operational tempo.

Artificial intelligence is reshaping the cognitive dimension of conflict, enabling faster decision-making, more granular intelligence fusion, and autonomous action across increasingly contested domains. Autonomy, especially in the form of uncrewed aerial, maritime, and land systems, is allowing states and non-state actors to project force with reduced risk to personnel, while expanding persistence and reach. Hypersonic delivery systems, meanwhile, challenge existing defence architectures through speed and manoeuvrability, compressing strategic warning timelines and introducing new complexities to deterrence and escalation management.

Furthermore, the return of great power competition is driving a more contested and multipolar strategic environment. States, non-state actors, and transnational networks are competing for influence not only in the physical domains of land, sea, air, and space, but across the digital and cognitive theatres that define modern security. This competition is not merely military. It encompasses economic systems, technological standards, regulatory influence, and informational control, as states seek to shape not just outcomes, but the rules and perceptions that govern global order.

Another major driver is the impact of environmental stress and resource scarcity.

Climate change is acting as an accelerant of instability, intensifying competition over water, food, energy, and arable land. The increasing frequency and severity of natural disasters, coupled with demographic pressures, are likely to drive internal displacement, state fragility, and geopolitical contestation in vulnerable regions. These pressures may generate new theatres of competition, particularly in regions where governance is weak and international influence is fragmented.

Recent conflicts have demonstrated how these forces combine in practice. The war in Ukraine has highlighted the operational impact of rapidly diffused technology. The widespread use of inexpensive drones, the tactical application of commercial satellite imagery, and the rapid repurposing of civilian tools for military use have all illustrated how the technological threshold for strategic disruption has lowered. Smaller and less industrially advanced actors are now capable of imposing high costs on more powerful states through asymmetric innovation and tactical agility. In this environment, the assumption that dominance can be secured through superior platforms alone is increasingly untenable. Success is being redefined by the ability to integrate, adapt, and act at speed, across domains and under pressure.

3.3 Conflict Acceleration and Diffusion

Modern conflicts increasingly unfold simultaneously across physical, digital, and informational domains, with artificial intelligence acting as a strategic multiplier. The June 2025 escalation between Iran and Israel reflects this shift: within days, both sides employed coordinated drone strikes, ballistic missiles, cyber operations, and disinformation campaigns, demonstrating how AI-enabled systems now amplify both the speed and complexity of escalation across multiple theatres at once.

As portrayed by several news outlets, the June 2025 Israel–Iran escalation showcased a strategic evolution in military engagement, driven by AI-enabled, low-cost autonomous systems. Israeli forces reportedly used AI-enhanced targeting, smuggled drones, and human intelligence to degrade Iranian air-defence radars and missile infrastructure, conducting pre-emptive strikes near Tehran and across multiple provinces ([AP News](#), [Euronews](#), [Military.com](#)). A separate [Euro News](#) article published on the 18th June 2025 states.



“Guided by spies and artificial intelligence (AI), the Israeli military unleashed a nighttime fusillade of warplanes and armed drones that it smuggled into Iran to quickly incapacitate many of its air defences and missile systems”.

Iran responded with a massive missile and drone barrage, launching hundreds of ballistic missiles and UAVs toward Israel and intercepting foreign aerial incursions. Coverage by [Al Jazeera](#), and [Reuters](#) confirmed both the scale of the attacks and Israel’s active air and missile-defence response.

These developments demonstrate how AI-driven autonomy, sensor fusion, and precision targeting are transforming escalation dynamics, reinforcing the need for sovereign control over AI systems, ensuring that targeting, strike coordination, and response decisions are explainable, legally accountable, and aligned with national strategic intent.

During the Iran–Israel of June 2025 escalation, Iran’s use of ballistic missiles highlighted the operational and technical challenges of missile interception. These high speed, high altitude projectiles are inherently difficult to track and neutralise, often exceeding the sustained capacity of even advanced air defence systems. Reports from U.S. and Israeli sources noted that layered defences were rapidly taxed by the volume and velocity of incoming threats, illustrating how saturation tactics can degrade even the most advanced intercept architectures ([WSJ](#))

The Iran–Israel conflict also exposed a critical vulnerability in the sustainability of defensive AI-enabled systems: the rate at which missile defence interceptors were

consumed. According to the same WSJ report, Israel’s air defence architecture faced extraordinary strain during the wave of Iranian missile and drone attacks.

This conflict also revealed a critical shift in the character of modern warfare: the use of AI-enabled, low-cost, self-flying drones to overwhelm traditional air defence systems. Unlike ballistic or cruise missiles, which follow pre-programmed trajectories and require costly precision manufacturing, these autonomous or semi-autonomous systems can navigate, swarm, and retarget dynamically, using AI to optimise flight paths, evade defences, and saturate adversary decision loops. This evolution, already visible in Eastern Europe and the Middle East conflicts, marks a fundamental transformation in the economics and tempo of attack, as adversaries can now flood defensive systems with inexpensive, adaptive platforms at scale.

While missile defence interceptors and sensor systems have kept pace in some cases, the asymmetry between low-cost offensive autonomy and high-cost defensive interception introduces new operational and strategic risks. AI-enabled defence systems responsible for sensor fusion, targeting prioritisation, and interceptor allocation must therefore operate within sovereign control, ensuring that automated responses remain explainable, legally accountable, and aligned with national strategic intent in the face of such massed, adaptive threats.

3.4 Implications for Capability Development

The globalisation of digital infrastructure has created structural dependencies that undermine operational independence. Many AI systems, especially those used in defence, are built on software frameworks and hardware components sourced through extended, internationalised supply chains.

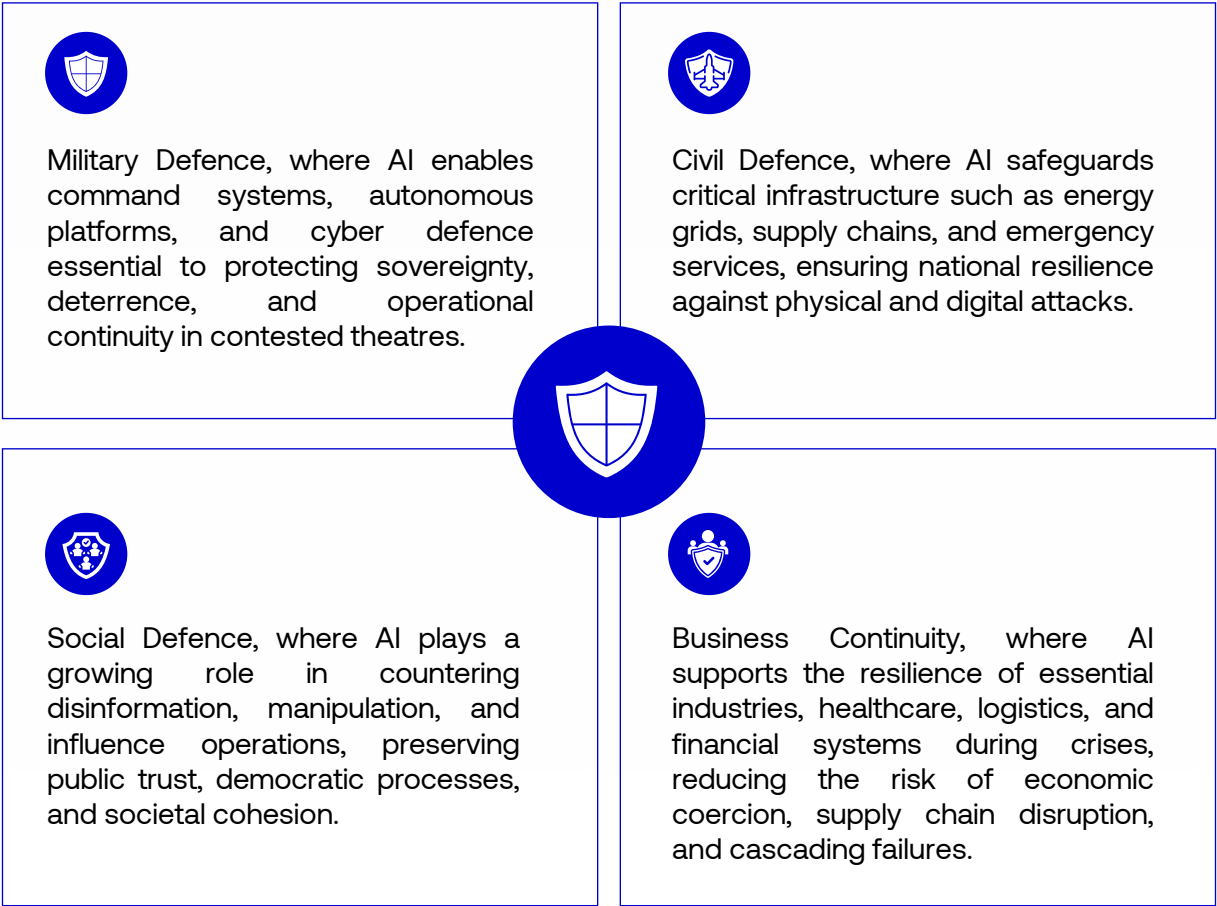
These include model weights, firmware, APIs, cloud compute, and silicon-level dependencies. As reliance on these systems deepens, so too does exposure to disruption, embargo, or unanticipated interference.

This not only challenges conventional forms of resilience, such as data integrity, network continuity, and supply chain security but also the resilience of decision-making itself. In crisis scenarios, the ability to generate trusted, context-sensitive, and legally accountable decisions must be preserved even under degraded conditions or adversarial manipulation. AI systems that operate as black boxes, depend on external update cycles, or cannot be validated in real time may erode that ability precisely when it is needed most. Sovereign AI is not simply about securing infrastructure; it is about safeguarding the integrity, explainability, and coherence of national decision-making under stress.

A more graduated approach is necessary, recognising that sovereignty is not a single threshold but a multidimensional framework that can be applied selectively based on mission criticality, operational risk, and legal

sensitivity. This approach acknowledges that a high degree of sovereign control is essential for high-consequence systems, such as targeting, command decision support, cyber defence, and battlefield autonomy, where external interference, misalignment, or denial could create unacceptable mission risk. In this context, sovereign control does not imply absolute control over every hardware or software component, but the assured ability to govern, validate, and if necessary, intervene in the behaviour of these systems.

AI is no longer confined to traditional military systems. Its accelerating integration across a broad range of national functions demands a more comprehensive understanding of the environments where sovereign AI control must be preserved. This paper identifies four interdependent security environments where the application of AI carries strategic, legal, and operational consequences:



While this paper focuses primarily on military applications, the principles of sovereign AI governance, assurance, and legal control must extend across this full spectrum of security environments. The ability to protect decision-making, critical infrastructure, and societal stability from digital coercion or loss of control is foundational to modern sovereignty.

Preserving sovereignty across these diverse security environments demands robust governance mechanisms that ensure AI systems remain under national control, oversight, and lawful intervention regardless of where or how they are deployed. The risks may differ, ranging from kinetic effects in military operations to economic disruption or cognitive manipulation, but the underlying need for trusted assurance remains constant.

Some AI applications in defence such as logistics optimisation, supply chain management, or personnel planning are often seen as commercially transferable, with limited direct impact on operational decision-making or the use of force. However, as recognised in the [EU AI Act](#) risk-based approach which states that the risk profile of an AI system is not determined by its technical function alone, but by its intended use, operational context, and potential impact on fundamental rights, safety, and mission outcomes. AI systems used for logistics, for example, may pose minimal risk in peacetime inventory management, but the same systems, if applied to battlefield supply chains or contested mobilisation, could carry significant operational risk or create points of

adversarial exploitation. Similarly, personnel analytics tools could inadvertently shape decisions with long-term human impact if not governed under clear legal and ethical frameworks. This reinforces the need for mission-driven sovereignty: where the degree of sovereign control is calibrated not by technical category, but by the assurance threshold appropriate to each function's role in defence operations.

Decentralisation of procurement practices has emerged as a key enabler of agility in AI deployment, the experience of Ukraine offers a salient example: under conditions of extreme operational urgency, the need for rapid acquisition and fielding of new technologies prompted a deliberate shift away from rigid, centralised procurement systems. By distributing procurement authority to lower command levels, Ukraine was able to accelerate innovation uptake and respond more dynamically to battlefield needs.

Granting more procurement autonomy to tactical or operational echelons can foster speed, experimentation, and local adaptation, especially in domains where commercial innovation cycles far outpace traditional defence timelines. Complementing this approach, several jurisdictions, including the United States, have begun to prioritise off-the-shelf solutions developed by small, local innovators over conventional defence primes. Companies like Palantir have exemplified this shift, offering modular, adaptable systems that integrate more fluidly with mission requirements while reducing time-to-field. In an April 2024 [blog post](#), Palantir stated.



“A key differentiator needed by the Department of Defense’s new systems, powered by emerging technologies, is enhanced modularity, openness, and flexibility. As software becomes increasingly central to achieving overmatch, we at Palantir believe that this kind of modular, open software will be a critical advantage that enables future ground, air, maritime, and space capabilities to achieve — and maintain — superiority.” Together, these trends point toward a procurement model better suited to the pace and complexity of AI-era competition.

3.5 Illustrative Examples from the Russia-Ukraine Conflict and Emerging Technologies



The war in Ukraine has served as a live testing ground for a range of new operational methods and technological innovations, many of which have significant implications for the future of warfare. One of the most prominent examples has been the **widespread and innovative use of drones** at all levels of the conflict. As stated in a 2024 [The Economist](#) article, “Killer drones pioneered in Ukraine are the weapons of the future. They are reshaping the balance between humans and technology in war”.

Ukrainian forces, often with minimal formal support, have adapted commercial drones such as DJI quadcopters for reconnaissance, artillery spotting, and direct attack roles using improvised munitions. These low-cost platforms have provided persistent situational awareness and allowed relatively lightly equipped units to target and destroy high-value Russian assets with surprising efficiency. Russia, initially slower to adapt, has increasingly responded with its own cheap drone swarms and electronic warfare systems designed to jam or spoof these UAVs. The intense “drone-versus-counter-drone” battle has highlighted how massed, expendable systems can neutralise traditional advantages in heavy armor and artillery. Ukraine’s tactical drones are “inflicting roughly two-thirds of Russian losses,” making them “twice as effective as every other weapon in the Ukrainian arsenal,” says a recent study by the [Royal United Services Institute](#).

A second critical innovation has been the

fusion of commercial satellite imagery and open-source intelligence (OSINT) with tactical military operations. Ukraine has leveraged partnerships with private satellite firms, such as Maxar and Planet Labs, to obtain near-real-time imagery of Russian force dispositions. These capabilities have dramatically increased the speed and granularity of situational awareness, enabling Ukrainian forces to anticipate and respond to enemy movements with unprecedented precision. However, these benefits have also revealed structural vulnerabilities. In a notable case, SpaceX’s Starlink service, crucial for Ukrainian battlefield communications, became a point of operational friction.

In September 2023, [Elon Musk](#) reportedly declined a Ukrainian request to extend Starlink coverage over Russian-occupied Crimea, fearing that such use could trigger escalation. The result was the failure of a planned Ukrainian naval drone strike near Sevastopol, as the drones lost connectivity mid-mission. This incident underscored a significant strategic liability: that unilateral decisions by private actors, outside the formal chains of military or governmental accountability, can directly constrain operational freedom of action during armed conflict. It illustrates the need for sovereign oversight not only over the technical integrity of AI and data systems, but also over their governance structures, including the contractual, jurisdictional, and political contexts in which they operate.

[Atlantic Council](#) reported in an article that by early March 2022, five commercial firms were sharing day and night satellite imagery that assisted Ukraine in tracking Russian forces. By December, Ukraine could tap into the “roughly 40 commercial satellites a day [that] pass over the area in a 24-hour period. Combined with extensive use of open social media data, Ukrainian forces have demonstrated a capacity for “crowdsourced intelligence” that has often outpaced traditional intelligence, surveillance, and reconnaissance (ISR) timelines.

This development foreshadows a future where almost any actor can access global surveillance tools previously available only to major powers. On May 7th 2025 major news outlets reported that China helped Pakistan to move satellites and recalibrate its air defence systems before it shot down Indian fighter jets ([Telegraph](#)).

Artificial intelligence has made its way onto the battlefield, though often in ways that remain below the threshold of full autonomy. AI has been employed in target recognition, data fusion, and predictive analysis to anticipate Russian movements and optimise the allocation of scarce resources. Ukrainian

and Western developers have used machine learning models to sift through vast quantities of drone footage, satellite imagery, and intercepted communications, helping commanders make faster and more informed decisions. Although still nascent, these AI-supported systems are pointing towards a future where decision cycles are compressed dramatically, and where human operators are increasingly assisted or even replaced by automated analysis at critical points in the command chain. [MIT Technology Review](#).

Finally, both sides have demonstrated the increasing importance of electronic warfare (EW) and cyber operations as integral parts of their campaigns. Russian forces have deployed significant EW assets to jam Ukrainian communications and GPS signals, while Ukraine has conducted successful cyber attacks against Russian logistical systems, disinformation platforms, and even critical infrastructure. The symbiosis of kinetic and non-kinetic attacks where, for example, a cyber attack disables a supply depot’s ordering system ahead of a precision strike is becoming more refined and coordinated.

3.6 Non-Compromising Interoperability

Sovereign AI must be designed in the context of a highly interdependent security environment. Defence operations are rarely conducted unilaterally. Most military actions occur within coalitions, alliances, or multilateral frameworks that require interoperable systems, shared situational

awareness, and coordinated decision-making. From an AI perspective, alliances such as NATO provide not only operational partnerships but also significant strategic benefits for sovereign capability development. These include:



Shared standards for AI safety, ethical use, and legal accountability, reducing fragmentation and promoting mutual trust in AI-enabled operations.



Pooling of mission-relevant datasets (including ISR, cyber threat intelligence, and synthetic training environments) that no single nation could generate alone.



Co-investment in foundational AI research, compute infrastructure, and capability validation, lowering the barriers to sovereign capability for smaller or resource-constrained allies.



Deterrence signalling by embedding AI into collective defence postures, states can benefit from allied assurance without surrendering decision authority.

That said, there is an inherent spectrum of dependency within any alliance. Total technological reliance on partner states introduces strategic liabilities, vulnerabilities to denial, manipulation, or political divergence in moments of crisis. The challenge is to strike the right balance: to build systems that are interoperable without being dependent, capable of joint operation without surrendering sovereign control.

This paper advances the principle of non-compromising interoperability: a model in which AI systems are designed to operate

together with allies, but in ways that preserve each nation's legal frameworks, assurance standards, and ultimate command authority. By adopting this posture, states can benefit from the shared strength of alliances while maintaining the freedom to act, decide, and govern under their own laws. In short, **non-compromising interoperability enables nations to operate "together, but sovereignly"** ensuring that cooperation enhances collective security while safeguarding each nation's freedom to act in its own interest under its own laws.

3.7 Expansion of Defence Concept: Blended Warfare Across Domains

The concept of defence has evolved beyond the traditional confines of armed forces and territorial sovereignty. In the contemporary environment, defence must be understood as a multidomain undertaking that spans military, civil, social, and commercial

domains. Modern conflicts do not occur solely on battlefields; they are fought across infrastructure networks, financial systems, information ecosystems, and the everyday digital experiences of populations.

Military defence remains foundational.

Conventional forces, deterrence capabilities, and operational readiness continue to play an essential role in safeguarding national security. However, the resilience of a nation's critical infrastructure, its energy grids, water supplies, transport systems, and communication networks is equally vital.

Civil defence

encompasses measures taken to protect the civilian population and infrastructure during times of war, natural disasters, or other emergencies. It ensures continuity by enabling societies to withstand both physical attack and digital disruption, maintaining core functionality during crises and enabling rapid recovery thereafter.

Societal defence has emerged as a critical pillar of national security

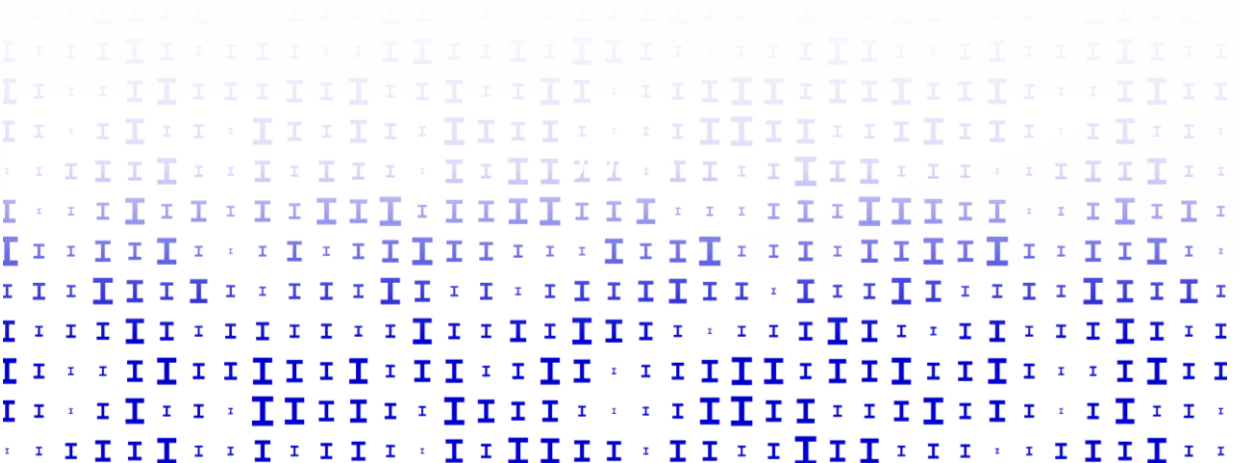
The information space is now a contested domain, where influence operations, disinformation campaigns, and narrative manipulation can achieve strategic effects without the need for kinetic action. Maintaining public trust, countering misinformation, and protecting democratic discourse are now as important to national resilience as securing physical borders.

Business continuity represents another indispensable element of modern defence.

Economic stability, industrial production, food security, medical supply chains, and access to critical technologies are all potential targets in blended conflict environments. Disruption of these systems whether through cyber attack, supply chain manipulation, or market coercion can weaken national resolve and strategic freedom of action without a single shot being fired.

In this context, Sovereign AI capabilities must be developed with an awareness that defence requirements extend far beyond the military. AI must support not only battlefield dominance but also the protection of civil infrastructure, the integrity of social discourse, and the resilience of critical

business functions. Governments must therefore adopt a holistic view of defence sovereignty. AI sovereignty policies must reflect this reality, ensuring that technological independence, operational assurance, and ethical governance are embedded across all domains of modern defence.



4. Rationale for Sovereign AI

Artificial Intelligence is not just a technological frontier; it is a contested domain of power, trust, and strategic judgement. As military systems become increasingly reliant on algorithmic decision support and machine driven operational functions, control over those systems becomes an issue of sovereignty, not merely capability. For the United Kingdom, which operates under legal and ethical constraints shaped by international humanitarian law, parliamentary oversight, and alliance coordination, the need to govern the AI systems that underpin military force is acute.

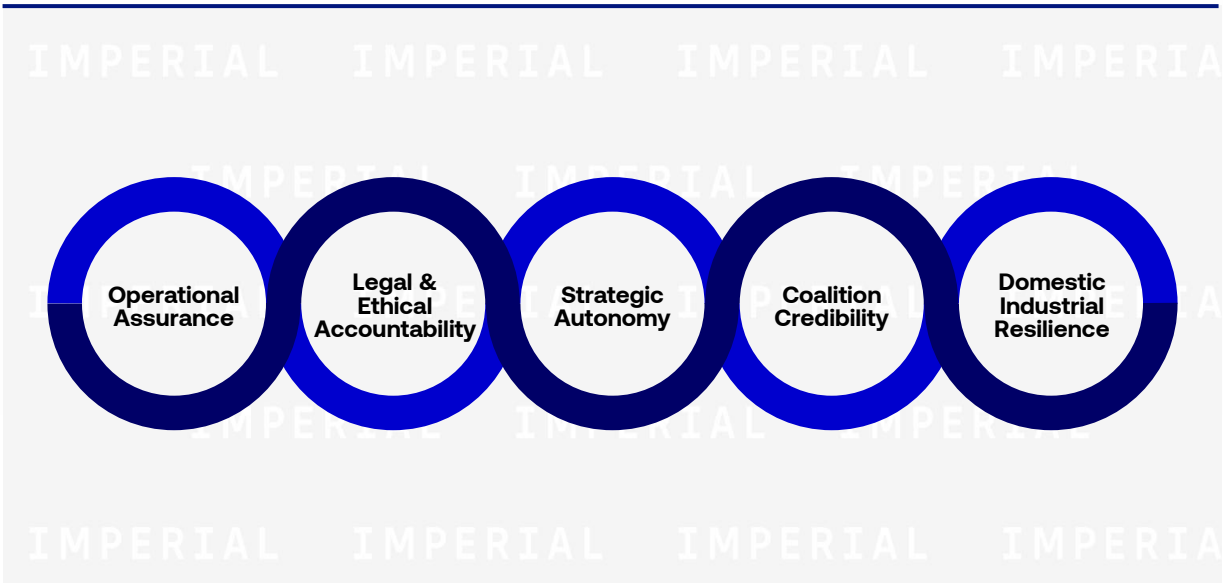
The [UK's Modern Industrial Strategy \(2025\)](#) sets a clear national direction for strengthening sovereign capability in critical technologies. It identifies Defence as one of eight priority sectors under its IS-8 growth framework, alongside Digital and Technologies, recognising that sovereignty in areas such as artificial intelligence, quantum computing, and semiconductors is central to both national security and economic resilience. The strategy commits over £500 million through the creation of a dedicated

Sovereign AI Unit, alongside £670 million to advance UK quantum computing, and the establishment of AI Growth Zones to accelerate adoption of advanced technologies across key industries.

These commitments are also reinforced by the [AI Opportunities and Action Plan \(2025\)](#), which sets out the UK's ambition to lead in trusted, secure, and Sovereign AI, while maintaining a globally competitive innovation ecosystem. The action states that the government should “Create a new unit, UK Sovereign AI, with the power to partner with the private sector to deliver the clear mandate of maximising the UK's stake in frontier AI”.

Both documents make clear that sustained investment in compute infrastructure, AI skills, and digital sovereignty is essential not only for industrial growth but for maintaining strategic freedom of action in defence, particularly in the development, governance, and deployment of AI-enabled military capabilities.

The strategic rationale for pursuing Sovereign AI in defence rests on five interdependent logics:



4.1 Operational Assurance and Mission Integrity

Modern defence operations require speed, agility, and coherence across multiple domains: land, sea, air, space, and cyber. As operational tempo increases and the volume of data surpasses human capacity for real-time assessment, AI becomes an indispensable force multiplier. However, this reliance introduces vulnerability if the MOD does not control how these AI systems are trained, updated, and deployed.

In time sensitive operations such as dynamic targeting, cyber incident response, or unmanned platform navigation, the capacity to override, retrain, or verify an AI model becomes mission critical. Should an AI-enabled system behave unpredictably, respond to spoofed input, or produce recommendations outside UK rules of engagement, the MOD must retain the authority and technical access to intervene. This is only possible when AI systems are sovereign at the point of use and at the level of behavioural governance.

Sovereignty in AI is often framed in terms of legal authority or technical access but both are hollow without the human capability to act on them. A state may possess the legal right to govern a system and the infrastructure to host it, but without a skilled cadre of technologists, engineers, operational analysts, and legal-auditors trained to interrogate and interpret AI behaviour, that sovereignty cannot be meaningfully exercised. **Sovereign AI therefore depends not only on control over systems, but on sustained investment in cleared, qualified, and strategically aligned personnel** who can adapt these systems in real time, test their outputs, retrain them to mission shifts, and

ensure they perform in accordance with both law and intent.

While AI may displace certain manual or routine defence roles, particularly in logistics, monitoring (peacetime), and procedural intelligence analysis, it will also generate new demand for highly skilled personnel in areas such as model assurance, red-teaming, legal-technical governance, and sovereign system integration. The centre of gravity for human involvement is shifting, from system operation to system stewardship. This reinforces the paper's core argument: that **Sovereign AI requires not fewer humans, but differently placed ones, those able to govern, adapt, and justify the behaviour of complex AI systems under national control.**

[MOD's Joint Doctrine Publication 04 \(JDP 04\)](#) emphasises the importance of achieving clarity in understanding complex operational environments. It notes that understanding is a continuous process that draws on critical thinking, judgement and assessment to make sense of complexity and ambiguity. The document cautions that flawed understanding can arise when individuals rely on incomplete or biased inputs, and stresses the central role of human judgement and reflection in making sense of dynamic information. In this context, as AI tools increasingly shape operational insight, the need to retain critical human oversight becomes paramount, ensuring that situational interpretation remains anchored in accountable, trusted processes. Sovereign AI enables operational assurance by preserving control over the models that shape perception and action.

4.2 Legal and Ethical Responsibility

Under international law, the use of force by state actors must meet stringent tests of necessity, proportionality, and distinction. These obligations extend not only to the act of force but to the means by which information is processed and decisions are made. AI systems used in targeting support, operational planning, or threat identification become part of the legal chain of accountability.

The House of Lords 2023 report on [AI in Weapon Systems](#), underscores the critical necessity of maintaining human control over autonomous weapon systems (AWS) throughout their entire lifecycle. The report emphasises that such control is essential to ensure compliance with international humanitarian law and to uphold ethical standards in military engagements. It highlights that the unpredictability and complexity inherent in AI technologies necessitate robust mechanisms for human oversight to prevent unintended actions and to maintain accountability.

Furthermore, the report calls for the UK Government to lead international efforts in establishing clear definitions and regulations concerning AWS. It stresses the importance of developing an international consensus on the criteria that AWS must meet to be considered compliant with legal and ethical standards. Central to this initiative is the retention of human moral agency in the decision-making processes of AWS, ensuring that machines do not have the ultimate authority in life-and-death situations. By advocating for these measures, the report aims to balance the potential operational advantages of AI in weapon systems with the imperative to uphold ethical standards, legal obligations, and public trust in military operations. If the MOD were to rely on AI systems trained abroad, hosted on foreign servers, or governed by proprietary logic unavailable to UK auditors, it would be unable to meet this legal threshold.

A critical dimension of national AI sovereignty is the capacity of the public sector to manage AI systems and respond effectively to the emerging threats they entail. This includes the need to train public sector employees in specific skills, particularly around cybersecurity, AI governance, and risk management. Human capital must be seen as a strategic asset in safeguarding digital sovereignty and institutional resilience. To build this capacity, governments must modernise the digital infrastructure of public institutions while simultaneously investing in local skills and training programs. This capacity building must be locally rooted to ensure sustainability, reduce dependency on foreign actors, and align with national security imperatives.

The risks of failing to act are tangible. In Costa Rica, a major [ransomware attack](#) in 2022 forced the government to temporarily shut down the computer systems used to declare taxes and for the control and management of imports and exports, causing an economic loss of about US\$ 125 million in the first 48 hours following the attack. Furthermore, teachers were unable to get paychecks, tax and customs systems were paralysed and health officials were unable to access medical records. On 8 May 2022, the president of Costa Rica issued an executive order proclaiming a national emergency due to the cyberattacks against the country's public sector and stated that the country was in a "state of war".

In parallel with technical capacity building, there is a pressing need to prepare the "human in the loop" for their evolving role in operational governance. As Sovereign AI systems become increasingly embedded in critical decision making processes, it is not enough for these systems to be merely auditable and traceable, they must also be fully understood by the individuals responsible for overseeing and ultimately authorising their outputs.

To achieve this, governments must invest in tailored training programs and clear operational guidelines that equip public sector employees with the knowledge and judgment required to manage human-machine interactions responsibly. A key area of focus is the “subconscious bridge”, the subtle psychological dynamic that shapes how humans interpret and respond to AI-generated recommendations. This includes understanding the legal and ethical liabilities that arise when human decisions are influenced by, or dependent on, algorithmic input.

Moreover, it is essential to address common cognitive biases that may compromise decision-making. For example, confirmation bias can lead individuals to accept AI outputs that reinforce their pre-existing beliefs without critical analysis. Similarly, machine bias, the tendency to over-trust a system that has performed reliably in the past, can result in the uncritical acceptance of AI

recommendations, even in cases where the system may hallucinate, propagate misinformation, or behave unpredictably.

Building this level of AI literacy is crucial to safeguarding against systemic overreliance and maintaining strategic human judgment in sovereign systems. **Public servants must be empowered to critically assess, override, or disengage from AI-driven decisions when risks or inconsistencies are identified.**

Finally, the public sector must go beyond initial implementation and build long-term institutional capabilities for the maintenance, auditing, and monitoring of AI systems. This includes establishing robust lifecycle tracking mechanisms, feedback loops, and ongoing performance and alignment evaluations to ensure that AI tools continue to operate in line with national objectives, legal standards, and public interest.

4.3 Strategic Autonomy and Freedom of Action

Continuing with the UK as our reference, the nation’s ability to act independently in defence of its interests relies on more than hardware. It depends on decision-making systems, intelligence fusion pipelines, cyber-defence architecture, and command coherence. **If the MOD becomes dependent on external entities for AI functionality, especially those bound by foreign legal systems or commercial incentives, it risks losing freedom of**

action in moments of crisis or divergence.

[The Integrated Review Refresh \(2023\)](#) emphasises the UK’s commitment to “maintain the UK’s freedom of action, freedom from coercion and our ability to cooperate with others” in an increasingly contested global environment. AI is at the heart of that technological contest. Sovereign AI is a shield for strategic autonomy.

4.4 Alliance Interoperability on Sovereign Terms

The UK is a core member of NATO, the Five Eyes intelligence partnership, AUKUS, and the Joint Expeditionary Force. These alliances depend on interoperability, shared threat models, and coordinated action. But

interoperability does not imply uniformity or dependence. Indeed, the most credible partners are those who can contribute sovereign capability under known and trusted parameters.

NATO's summary on [Artificial Intelligence Strategy \(2021\)](#) states that "Artificial Intelligence (AI) is changing the global defence and security environment. It offers an unprecedented opportunity to strengthen our technological edge but will also escalate

the speed of the threats we face. This foundational technology will likely affect the full spectrum of activities undertaken by the Alliance in support of its three core tasks; collective defence, crisis management, and cooperative security."

Allies and [NATO](#) commit to ensuring that the AI applications they develop and consider for deployment will be in accordance with the following six principles:

Lawfulness

AI applications will be developed and used in accordance with national and international law, including international humanitarian law and human rights law, as applicable.

Responsibility and Accountability

AI applications will be developed and used with appropriate levels of judgment and care; clear human responsibility shall apply in order to ensure accountability.

Explainability and Traceability

AI applications will be appropriately understandable and transparent, including through the use of review methodologies, sources, and procedures. This includes verification, assessment and validation mechanisms at either a NATO and/or national level.

Reliability

AI applications will have explicit, well-defined use cases. The safety, security, and robustness of such capabilities will be subject to testing and assurance within those use cases across their entire life cycle, including through established NATO and/or national certification procedures.

Governability

AI applications will be developed and used according to their intended functions and will allow for: appropriate human-machine interaction; the ability to detect and avoid unintended consequences; and the ability to take steps, such as disengagement or deactivation of systems, when such systems demonstrate unintended behaviour.

Bias Mitigation

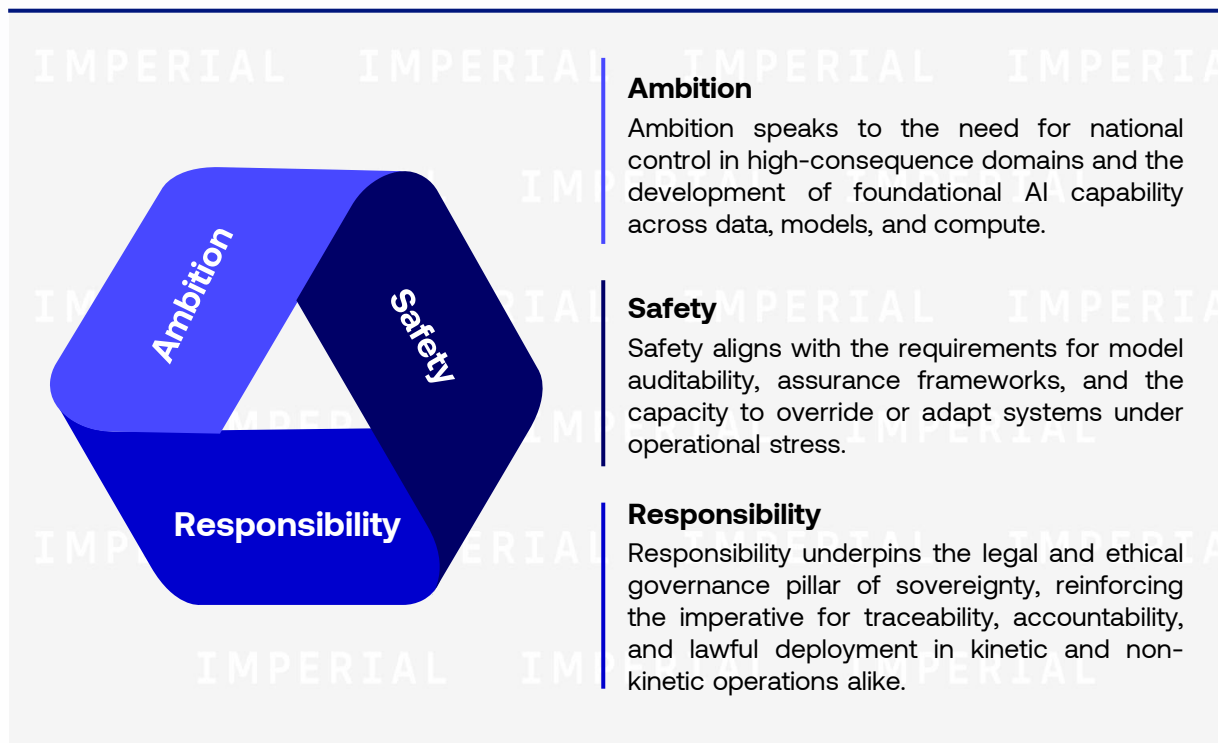
Proactive steps will be taken to minimise any unintended bias in the development and use of AI applications and in data sets.

This implies a requirement for national oversight, even within a multilateral framework. The UK's contribution to AI enabled joint operations will be most valuable when its systems are dependable, secure, and independently verified, not borrowed or externally dictated.

Similarly, The MOD's Ambitious, Safe, Responsible (ASR) policy statement on AI sets out the principles through which

Defence will govern the development, deployment, and use of AI. The document emphasises that the UK will lead by example in ensuring that AI is developed, deployed, and governed responsibly, safely, and transparently. It further outlines the need for clear accountability structures, robust legal oversight, and a strong ethical foundation for AI in defence, aligning closely with NATO's principles while reinforcing the UK's national priorities.

These principles map directly onto the dimensions of sovereignty articulated in this paper.



Together, they offer a coherent doctrinal baseline through which sovereignty can be operationalised within both national and alliance contexts. Sovereign AI enhances alliance cohesion by allowing the UK to contribute validated capabilities without fear

of data exposure, doctrinal misalignment, or legal incompatibility. It also allows the UK to shape the emerging standards for AI in military operations from a position of operational credibility.

4.5 Domestic Industrial and Economic Security

Sovereign AI is a critical instrument of national industrial strategy, but it should not be treated as a singular or absolute requirement. As discussed in earlier chapters, the degree of sovereign control must be calibrated to reflect the criticality, operational context, and enduring strategic value of each capability.

The UK [MOD's Defence and Security Industrial Strategy \(2021\)](#) supports this by stating that the government must preserve operational independence through the capability to design, build, and support

critical defence systems onshore. AI clearly falls within this remit, but this does not imply that every AI system must be developed and maintained exclusively within national borders.

The goal must not be to achieve total self-sufficiency across every layer of the stack, but to establish a robust and defensible capability that allows the state to operate critical AI systems under its own governance, particularly in scenarios where alliance interoperability is unavailable or contested. This requires a layered approach:



The emphasis is on a posture designed to retain mission continuity and decision authority even under adversarial or politically constrained conditions.

Investment in Sovereign AI capability strengthens the domestic innovation ecosystem by supporting the UK's SME base, stimulating long-term talent retention in high-end digital sectors, and ensuring that intellectual property generated for defence remains adaptable and exportable on national terms. It also mitigates the strategic and fiscal risks associated with vendor lock-in, opaque procurement pipelines, proprietary data formats, and externally mandated upgrade cycles. Far from conflicting with open collaboration, these

safeguards promote transparent standards, interoperable frameworks, and a more competitive supplier landscape, rather than closed monopolies.

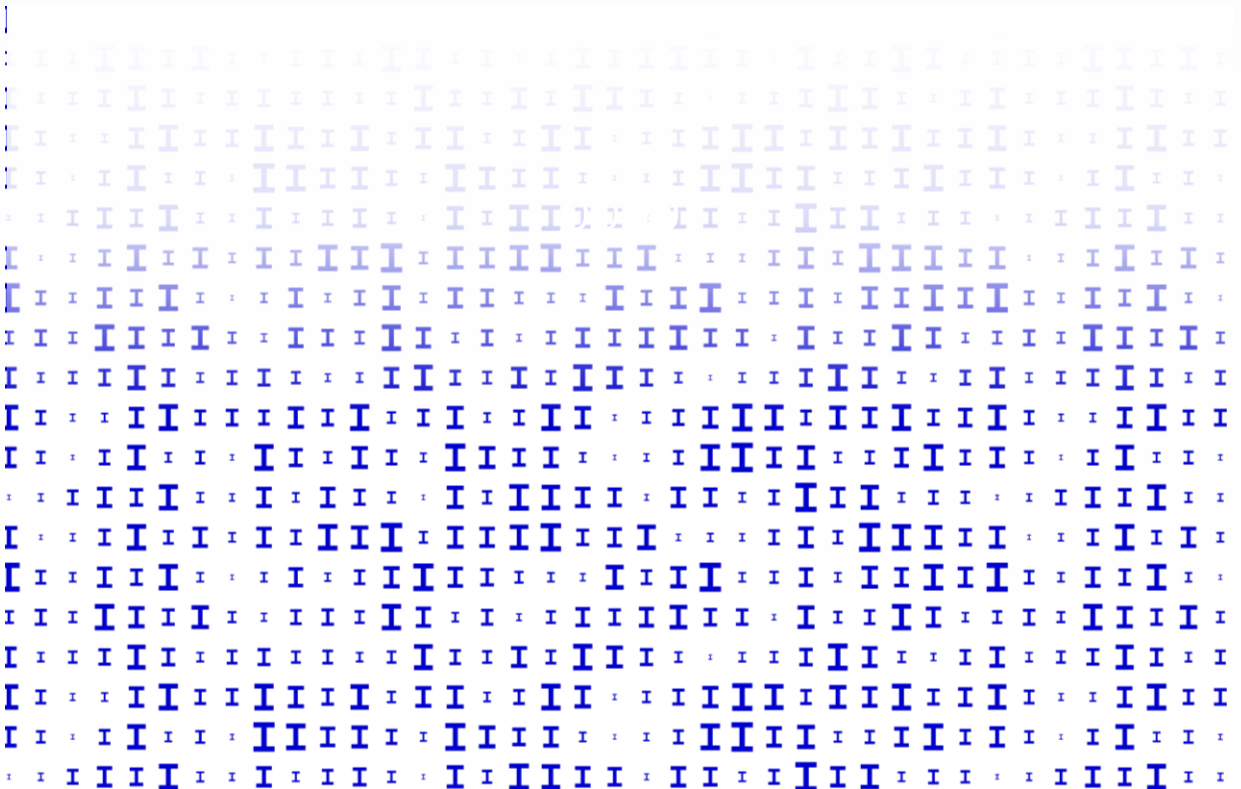
As AI assumes more cognitive and interpretive functions within defence workflows, its governance cannot be treated like conventional equipment. The distinction is closer to that between a state-trained military professional and a private contractor: one operates under national authority and institutional ethos; the other under a commercial logic, even if contractually aligned. **Sovereign AI ensures that military judgement, when delegated to machines, remains structurally embedded within the national chain of command.**

However, for lower-risk domains (in peacetime conditions and as recognised in the EU AI Act risk-based approach we covered earlier), such as logistics, personnel management, and routine data analytics, a more flexible approach may be appropriate. In these areas, sovereignty can be maintained not through direct control over every component, but through robust procurement oversight, contractual governance, and technical due diligence. This allows for the integration of commercial AI systems, provided they meet strict security standards, are properly validated, and are capable of integration without compromising core national interests. It also reflects the reality that demanding full control over every aspect of AI capability can limit access to cutting-edge innovations, reduce operational agility, and drive up long-term costs.

The AI-enabled military of the future will not simply purchase software; it will co-develop, integrate, and refine AI systems across the

lifecycle of operational capability. This co-development may involve public-private partnerships, particularly with domestic firms or international partners operating within trusted governance frameworks. Sovereignty ensures that this process serves national interests, but it does not require exclusive national ownership of every component, only that critical systems, data, and decision logic remain under assured, auditable, and legally accountable national control. Where co-development involves foreign entities, it must be governed through legal safeguards, data control provisions, and assurance regimes that prevent extraterritorial dependency or strategic compromise.

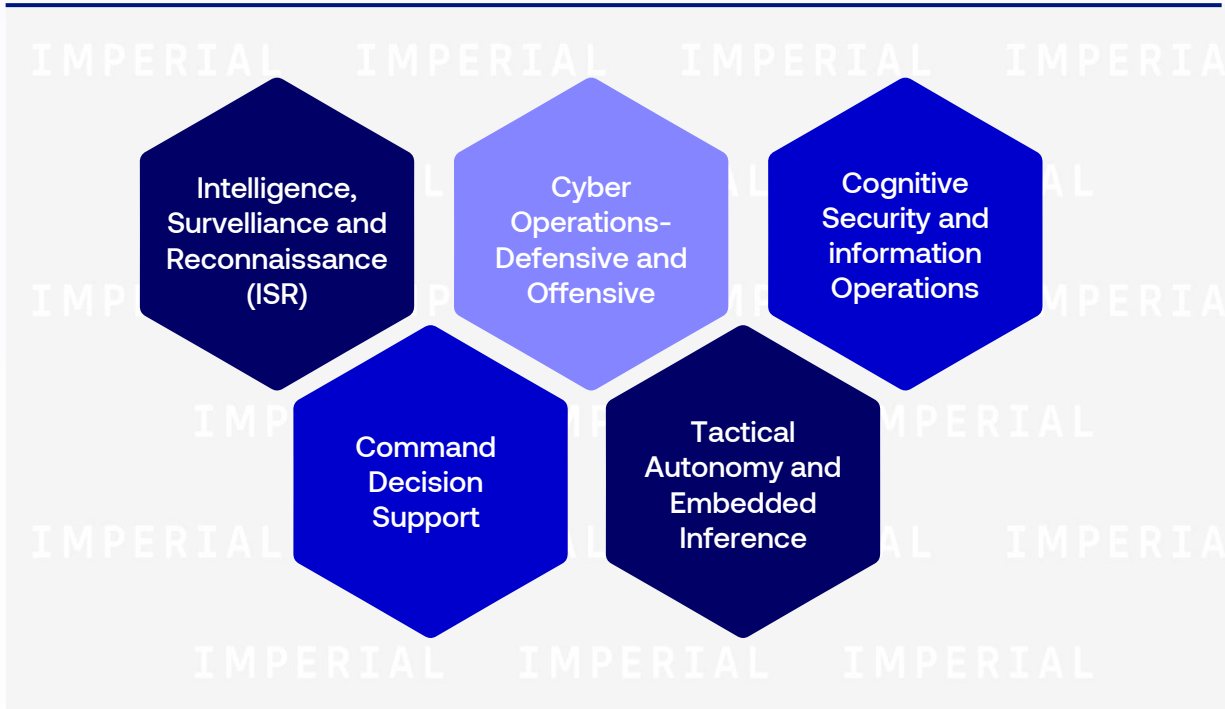
This approach prioritises national control where it matters most, while leveraging commercial innovations and international partnerships where appropriate. It not only supports the UK's strategic autonomy, but also its broader economic resilience, industrial base, and technological leadership in the digital age.



5. Operational Domains Requiring Sovereign AI

This chapter presents five mission domains as illustrative examples where Sovereign AI is likely to be most consequential. These areas are not exhaustive but reflect operational functions where the absence of sovereign

control could pose immediate or cascading risks. Actual prioritisation should be driven by strategic assessment frameworks and adaptive force design processes already embedded in defence governance.



5.1 Intelligence, Surveillance and Reconnaissance (ISR)

Modern ISR systems generate vast volumes of data, from high resolution satellite imagery and full motion video to electromagnetic spectrum emissions and behavioural analytics. Artificial intelligence is increasingly deployed to fuse, filter, and prioritise this data in real time, generating actionable insights to support force protection, threat assessment, and time-sensitive targeting. AI models deployed in ISR environments often use pattern recognition, sensor fusion, and object classification techniques to identify adversary movements, equipment signatures, and behavioural anomalies. In

doing so, they serve as operational proxies for judgment, flagging threats, validating targets, and sometimes influencing lethal decisions. The legal and strategic stakes are high.

If the MOD cannot trace how such models generate outputs or cannot control their training data, inference thresholds, or operational updates, it will be unable to ensure compliance with the principles of necessity, humanity, proportionality, and distinction under the Law of Armed Conflict (LOAC).

Four key principles underpin LOAC:

a) Military necessity

a state engaged in an armed conflict is permitted to use only that degree and kind of force, not otherwise prohibited by the law of armed conflict, that is required in order to achieve the legitimate purpose of the conflict". This principle contains four basic elements:

- I. the force used can be and is being controlled;
- II. since military necessity permits the use of force only if it is 'not otherwise prohibited by the law of armed conflict', necessity cannot excuse a departure from that law;
- III. the use of force in ways which are not otherwise prohibited is legitimate if it is necessary to achieve, as quickly as possible, the complete or partial submission of the enemy;
- IV. conversely, the use of force which is not necessary is unlawful, since it involves wanton killing or destruction.

b) Humanity

forbids the infliction of suffering, injury or destruction not actually necessary for the accomplishment of legitimate military purposes.

c) Distinction

separates combatants from non-combatants and legitimate military targets from civilian objects.

d) Proportionality

requires that the losses resulting from a military action should not be excessive in relation to the expected military advantage.

Furthermore, if the models rely on foreign trained architectures or cloud hosted inference, the MOD could lose control at precisely the moment decisive judgment is needed.

The use of AI-enabled weapon systems (AWS) raises fundamental questions about human oversight, legal accountability, and operational reliability. The House of Lords AI in Weapon Systems Committee report, [Proceed with Caution: Artificial Intelligence in Weapon Systems](#), underscores that human decision-making is central to legal accountability in the use of autonomous systems. It emphasises that accountability cannot be transferred to machines, and that meaningful human control must be integrated into all AI-enabled AWS to ensure clear human accountability on the battlefield.

In its response to this report, [The Government Response](#) has reinforced this position, agreeing that weapon systems must be used in a manner which is compliant with International Humanitarian Law (IHL). The Government's approach includes a layered governance framework that provides human oversight before, during, and after the deployment of AI-enabled military systems. This framework includes policy frameworks, risk management processes, system test and evaluation, operator training, targeting processes, parameter setting, battlespace management, and postoperative reporting and investigation.

Additionally, [Article 36 Weapons Reviews](#) play a critical role by requiring States to determine whether new weapons, means or methods of warfare may be employed lawfully under International Law. These reviews are not simple 'review and release'

events, but rather an iterative process that includes ongoing assessment as systems evolve, particularly as learning systems present new legal and operational challenges. This approach reflects the UK's commitment to ensuring that commanders and operators maintain full awareness of the capabilities and limitations of the systems under their authority, and that they retain the ability to exercise meaningful human control in compliance with international law.

The MOD has also committed to maintaining transparency with Parliament and the public regarding the governance processes for AI-enabled military systems. This includes ongoing engagement with expert stakeholders and international partners to share lessons, best practices, and insights on the safe, lawful, and ethical use of AI in defence. This collaborative approach is essential for addressing the unique challenges posed by emerging AI technologies and ensuring that IHL compliance remains robust as AI capabilities continue to advance.

In this context, the requirement for Sovereign AI in Intelligence, Surveillance, and Reconnaissance (ISR) is not simply a technological preference, but a legal obligation and an operational necessity. It ensures that decisions made by autonomous systems remain accountable, transparent, and legally defensible, reinforcing both operational credibility and strategic trust in the use of military AI. These systems generate sensitive, often siloed data streams, including full-motion video, SIGINT, and electronic emissions that, when governed under Sovereign AI frameworks, form part of the UK's underleveraged "dark data" ecosystem.

5.2 Command Decision Support

Machine learning models are now routinely employed to simulate adversary responses, optimise logistics flows, assess operational risk, and generate time-sensitive courses of action. These systems are increasingly embedded within joint planning

environments and tactical operations centres, where their outputs are used not merely as recommendations but as framing devices that structure human judgement under pressure.

When AI systems propose actions or project outcomes, they do more than support decision-makers, they define the contours of decision space itself. As noted in the [2022 Defence Artificial Intelligence Strategy](#), this creates a need for human centred AI that preserves meaningful **human control, ensures transparency of inference, and embeds legal accountability into system behaviour**. Without such safeguards, especially where sovereign control is lacking, there is a real risk that national military judgement could become subordinate to the logics of externally governed, opaque, or misaligned AI systems.

This risk is especially acute in scenarios where AI is employed in mission planning, threat anticipation, or nuclear posture modelling. In such high consequence applications, models must be developed, aligned, and governed entirely within the sovereign control perimeter of the state, with documented assurance of data provenance, inference logic, re-trainability, and override capability. Delegating these functions to non sovereign platforms introduces unacceptable uncertainty into strategic calculus.

Critically, **AI systems supporting command decisions do not operate in isolation**. As

seen in the war in Ukraine, logistics and supply chains, from fuel distribution and ammunition routing to depot resupply and force projection are not ancillary functions, but primary determinants of combat viability. **AI-enabled logistics is now tightly coupled with ISR, targeting, and manoeuvre planning**. In such environments, AI agents coordinate across shared datasets, interact through common command architectures, and adapt dynamically to both friendly and adversarial activity.

This interdependence presents a structural challenge to sovereignty. The assumption that AI systems can be cleanly partitioned by function or security classification is often untrue in practice. A vulnerability in one system, even if ostensibly isolated, may propagate through feedback loops, real-time data sharing, or unintended interactions, resulting in failure cascades or degraded decision assurance. In increasingly connected battlespace architectures, sovereignty must be understood not solely at the level of components, but across multi-agent units, where distributed inference and adaptive behaviour are the norm. The ability to validate, audit, and intervene in such systems requires governance mechanisms that span across systems and domains, not just within them.

5.3 Cyber Operations - Defensive and Offensive

In defensive cyber operations, AI is used to detect anomalies, classify threats, and automate incident response. These systems ingest and analyse privileged telemetry, including network logs, communications metadata, and low-level system events. They are central to the security of military infrastructure, classified communications, and deployed platforms. If the models supporting these functions are trained

externally, updated through unverified supply chains, or reliant on foreign platforms, they introduce unacceptable risks such as backdoors, misclassification, data leakage, or latent compromise.

The UK's National Cyber Strategy 2022 highlights the strategic imperative of supply chain integrity, stating:



“We will reduce our reliance on individual suppliers or technologies which are developed under regimes that do not share our values.”

Sovereign AI in this context must include control over the full model lifecycle, from data collection and training, to red-teaming, deployment, and rollback. Models must be resilient to adversarial manipulation, capable of being quarantined or overridden under contested conditions, and governed under nationally defined threat taxonomies and doctrinal logic. Red-teaming exercises, conducted under sovereign authority, must simulate adversarial attacks to test and validate detection efficacy and operational containment.

In offensive cyber operations, the risks and the need for sovereign control are even more pronounced. AI systems are increasingly used to support offensive tooling, including reconnaissance automation, vulnerability mapping, exploit generation, target environment modelling, and payload deployment planning. These capabilities lie at the intersection of technical execution and political signalling. They carry direct implications for deterrence posture,

escalation thresholds, and alliance integrity. The telemetry, network logs, and anomaly patterns produced across UK defence infrastructure represent another class of mission-specific data.

In both defensive and offensive contexts, cyber operations reveal the indivisibility of Sovereign AI. They show that control over data, model behaviour, and deployment infrastructure is not a theoretical abstraction, it is a practical requirement for the lawful, credible, and effective use of digital power. This domain is also the most immediate operational example of why sovereignty cannot be applied selectively or assumed retroactively.

States must therefore ensure that AI systems supporting cyber operations are fully sovereign across all six governance dimensions. In this domain, the ability to detect and act is inseparable from the ability to command and justify. Without Sovereign AI, both are compromised.

5.4 Tactical Autonomy and Embedded Inference

The deployment of autonomous systems across air, land, and maritime domains is increasingly central to modern military operations, particularly for high-risk or persistent missions where minimising human exposure is paramount. These platforms, ranging from unmanned aerial vehicles (UAVs) to autonomous ground and maritime systems, often rely on onboard inference engines to make rapid, low-latency decisions without continuous communication links.

The accelerating development of AI-enabled autonomy is transforming the character of tactical operations. In June 2025, China unveiled a [mosquito-sized drone](#) designed for stealth military operations, capable of penetrating air defences, conducting real-time surveillance, and potentially delivering targeted effects. These micro-autonomous

systems rely on embedded AI capable of processing real-time sensor feeds locally, enabling autonomous navigation, target detection, and decision-making without reliance on continuous human control or external communications.

Such developments point to a future where low-cost, swarming autonomous systems can be deployed at scale, dramatically increasing the speed, density, and complexity of the battlespace. Ensuring sovereign control over these systems means embedding failsafe behaviours, pre-programmed legal constraints, and robust decision boundaries directly into the deployed platforms, alongside the capacity for national authorities to retrain and validate models as threat environments evolve.

A notable example is the joint military exercise conducted by the United Kingdom, United States, and Australia under the [AUKUS](#) partnership. In this exercise, AI-powered autonomous drones were employed to detect and engage enemy vehicles, demonstrating the capability of these systems to operate cohesively and share data seamlessly to enhance response times and targeting accuracy. ([The Times](#)).

The integration of artificial intelligence into these systems effectively extends the Ministry of Defence's operational authority to deployed platforms and frontline systems, enabling decision-making to occur in real time and closer to the point of action. However, if the inference logic of these AI systems is opaque, unmodifiable, or cannot be overridden, it poses a risk to command integrity and accountability. The level and nature of human oversight must reflect the consequences, risk profile, and speed of decision-making inherent in the system's role. Tactical autonomy must be designed to operate in communications-denied or contested electromagnetic environments. The UK Army's [Robotics and Autonomous Systems \(RAS\) Strategy 2022](#) emphasises that RAS capabilities must be functional even where GPS or communications are degraded, enabling systems to "see, shift or shoot" across the battlefield while maintaining reliability and lethality stating that *"This hardware will have varying degrees of autonomy but never at the expense of meaningful human control."*

A complementary analysis [from Army University Press](#) reinforces this, stating: "AI-RAS are the solution to executing combat operations in a disrupted, degraded, or

denied GPS or communications environment. AI-RAS are more lethal. AI-RAS are more efficient. AI-RAS do not fatigue. AI-RAS are faster, stronger, more intelligent, and more rational than humans."

Sovereignty in the context of tactical autonomy does not require that every subsystem be designed and manufactured domestically. Systems procured from trusted allies such as the United States may still be integrated into sovereign force structures, provided they meet the conditions for sovereign governance. These include the ability to inspect and audit the AI logic, modify system behaviour to reflect national doctrine, and override or disengage automation under operational or legal review. In high-consequence applications, sovereignty is less about origin and more about control, accountability, and operational independence. The core requirement is that these systems operate within national command architectures, under rules of engagement defined by the procuring state, and with transparent pathways for validation, assurance, and fail-safe disengagement. Where such conditions cannot be met, whether due to black-box inference, update dependencies, or legal opacity the system must be classified as non-sovereign, regardless of alliance status. Tactical autonomy highlights the practical demands of Sovereign AI: not isolationism, but institutional capability to govern the behaviour of autonomous systems under contested, real-time conditions. **Sovereign AI in this domain enables lawful and effective force projection while preserving the integrity of national decision-making** in environments where machine logic and human judgement converge.

5.5 Cognitive Security and Information Operations

Cognitive security is now a central pillar of modern defence, encompassing the detection, disruption, and countering of adversarial information campaigns, influence operations, and digital propaganda. In this domain, artificial intelligence is used to monitor narrative environments, classify

disinformation, and respond to coordinated manipulation across open-source and classified channels. These tools increasingly inform decisions on how to respond publicly, diplomatically, or operationally to grey-zone actions below the threshold of armed conflict.

Yet the case for Sovereign AI in this domain is not always obvious. Many of the tools used for social media monitoring, language modelling, or semantic analysis are commercially available and widely deployed. However, it is precisely this dependence on externally governed systems, particularly those developed by large language model providers, content moderation platforms, or cloud-based NLP services that creates a critical vulnerability.

The risk is twofold. First, these systems are trained on publicly available data, often with embedded biases, misaligned incentives, or moderation policies shaped by commercial or foreign legal norms. This means their classifications of “malicious content,” “harmful speech,” or “coordinated activity” may not align with national definitions of threat, democratic standards, or legal evidence thresholds. Second, because the behaviour of such models is largely uninspectable, states cannot be confident in how responses are generated, how threat attribution is reached, or whether adversarial manipulation has already influenced the model’s behaviour. Without sovereign control, states risk outsourcing decisions about what constitutes a threat, who is responsible, and how to respond to systems they cannot audit, influence, or explain. This is not simply an operational problem. It is a strategic liability. In the context of information warfare, credibility, legitimacy, and escalation control rest on a state’s ability to defend not only its territory, but its narrative.

Sovereign AI in cognitive operations ensures that detection pipelines, narrative triage systems, and content response models reflect national legal frameworks, strategic priorities, and ethical boundaries. It allows governments to act with confidence, knowing that the tools used to assess information threats are aligned with domestic law and not subject to arbitrary

opaque training histories.

Sovereign AI in cognitive security is not about controlling speech or policing discourse. It is about ensuring that decisions about malign influence, escalation signalling, or digital sovereignty are made through tools that the state owns, governs, and can defend. In information warfare, credibility begins with control. The National Security and Online Information Team (NSOIT), formerly known as the Counter Disinformation Unit, leads the UK’s governmental response to misinformation and disinformation. NSOIT analyses publicly available information to identify and counter false narratives that threaten national security.

The proliferation of AI-generated deepfakes presents a growing challenge, with projections indicating a significant increase in such content. The Accelerated Capability Environment (ACE), a Home Office innovation unit that brings together experts from government, industry, and academia to rapidly prototype and deliver digital solutions for national security, has underscored the growing challenge posed by AI-generated deep fakes. In a government case study, ACE supported the development of tools to detect synthetic media, highlighting the urgent need for effective technological responses to counter digital impersonation and disinformation threats.

Sovereign AI systems play a crucial role in defending democratic discourse, ensuring that tools used for this purpose are governed by national ethics and not outsourced to platforms with opaque accountability or divergent political commitments. By maintaining control over AI systems and their underlying algorithms, a nation can uphold the integrity of its information environment and safeguard its democratic institutions.



6. Economic Modelling and Feasibility

Sovereign Artificial Intelligence is not an abstract ambition. It is a material question of capability, infrastructure, and cost. The decision to develop Sovereign AI within the UK Ministry of Defence carries with it a tangible set of economic implications, which must be addressed with the same seriousness as any major weapons system, infrastructure programme, or force development initiative. As this white paper has argued, sovereignty in AI is not optional in domains that implicate the legal, ethical, or strategic core of British defence. But it is equally true that this sovereignty must be financed, staffed, built, and sustained. A credible Sovereign AI posture must therefore demonstrate not only strategic logic, but economic viability and industrial realism.

At the core of this feasibility analysis is the

recognition that Artificial Intelligence at scale is resource intensive. Models suitable for critical defence functions such as targeting support, threat prediction, or cyber anomaly detection require advanced compute infrastructure, secure data environments, and skilled personnel. However, unlike commercial frontier AI models, such as large language models designed for general-purpose deployment across billions of users, **defence AI models are typically smaller, more focused, and tuned to specific missions.** As such, they can be delivered under a more contained economic envelope, provided they are developed with clearly defined objectives and integrated with existing digital assets across Defence Digital, DSTL, and allied research and procurement programmes.

While recent attention has focused on large language models (LLMs), defence applications rely on a broader range of AI techniques. These include:



A credible sovereignty posture must account for this diversity, ensuring control and assurance mechanisms are applied not only

to language-based models, but across the full spectrum of AI methods critical to defence operations.

6.1 Infrastructure Requirements for Sovereign AI

Developing sovereign artificial intelligence models suitable for critical defence functions such as ISR fusion, command decision support, and embedded autonomy demands significant infrastructure investment. Based on detailed projections in the [Considerations](#)

[Regarding Sovereign AI and National AI Policy](#), by the Trusted AI Alliance at Imperial College London, establishing the AI training and hosting infrastructure required for sovereign capability within the UK is technically feasible, though capital-intensive.

To achieve the training of a model on the scale of 1 trillion parameters, a future facing benchmark that enables generalisation across multi domain defence missions the infrastructure is expected to include:

Capability	Minimum Parameters
Basic Language Understanding	1.5 billion (GPT-2)
Translation	10 billion
Coding	50 billion
Common Sense Reasoning	100 billion
Zero-shot Learning	175 billion (GPT-3)
Advanced Question Answering	500 billion
Complex Problem Solving	1+ trillion (GPT-4)

Source: Trusted AI Alliance, Imperial College London

There are limited published data points to inform these estimates; however, existing benchmarks such as the training of GPT-3 with 175 billion parameters (latest versions have not been officially confirmed by Open AI but experts believe GPT-4 uses approximately of [1.8 trillion](#) parameters) already demonstrate the extensive computational resources required for frontier models. Assuming systematic linear scaling in computational demand, training a 1 trillion-parameter model would necessitate approximately 10,000 NVIDIA H200 GPUs to achieve completion within operationally feasible timeframes. This projection accounts not only for model size but also for the increasing complexity of training tasks and the cumulative lessons learned from four major foundational commercial models.

The proposed hardware specifications for a UK Sovereign AI system, modelled on GPT-class architectures are driven by the operational need to support a general purpose language model of this scale. A model with approximately 1 trillion parameters is viewed as essential to deliver advanced capabilities such as natural language understanding, multi-modal reasoning, adaptive code generation, and other forms of strategic cognition. This

requirement establishes a clear pathway for defining the necessary compute architecture and infrastructure footprint, as well as estimating both capital expenditure and ongoing operational costs.

Beyond the technical rationale, investment in Sovereign AI infrastructure provides strategic benefits: it ensures national control over critical defence-relevant technologies, reduces exposure to foreign platform dependencies, and strengthens the resilience of the UK's digital and command systems. Additionally, a sovereign compute estate allows for modular retraining and capability refresh cycles, supporting long-term adaptability and cross-domain application integration across Defence.

The long-term trajectory of AI sovereignty will not be limited to control over software models, data pipelines, or cloud compute. As the field advances, AI systems are increasingly being co-designed with domain specific hardware, optimised not only for speed but for safety, locality, and environmental constraints. This marks a critical evolution: **compute is no longer neutral infrastructure, but a shaping layer of algorithmic behaviour, security affordances, and deployment feasibility.**

Sovereign compute infrastructure must be understood not only in terms of processing capacity, but as an integrated capability spanning trusted hardware, secure energy supply, and resilient hosting architecture. It requires more than just access to GPUs, it depends on grid-secure power provision, trusted silicon supply chains, and scalable hosting environments capable of operating across multiple availability zones.

In alignment with the MOD's Defence AI Strategy and the British Army's mission to become "AI-Ready" across all force elements, **Sovereign AI capabilities should prioritise smaller, optimised models deployable at the tactical edge.** According to the British Army's official [Approach to AI \(Oct 2023\)](#), these systems must be scalable "from back office to battlefield" and tailored

to operate on ruggedised, low-power processors embedded within vehicles, drones, and dismounted kits. *"The weaponisation of data – in both the physical and virtual domains requires tangential thinking, to ensure we maintain pace with the high velocity technology changes associated with machine intelligent processes."*

To ensure continuity under degraded or contested conditions, federated compute models should be prioritised, enabling distributed inference, retraining, and rollback across secure domains. Institutional responsibility must be clearly defined, for example, in the UK, this could follow a structured model with distributed leadership across key authorities.

The MOD Strategic Command and Defence Digital leads infrastructure planning and operational deployment

DSIT coordinates national policy on secure silicon access and digital resilience

Crown Hosting and the Cabinet Office Digital function can play enabling roles in provisioning, security auditing, and scaling sovereign digital infrastructure across government and defence.

The UK's traditionally open market stance toward AI-relevant companies, such as the sale of DeepMind to Google and ARM to SoftBank has yielded global prominence in research and commercial innovation. However, this liberal acquisition environment has also reduced the state's strategic grip over foundational assets in compute design, model development, and platform engineering. As AI becomes more deeply linked to national security, industrial policy must evolve to reflect the criticality of certain capabilities. The pursuit of Sovereign AI may require a reassessment of investment protections, strategic acquisitions, or targeted state participation in firms developing models, chips, or enabling infrastructure with defence applications. Open market innovation and sovereign

resilience are not inherently incompatible but they must be balanced against national risk exposure.

More broadly, AI must be understood not only as a discrete technological sector, but as a cross-cutting enabler of sovereign science and engineering capability. It accelerates materials discovery, systems design, logistics simulation, and threat modelling. As such, investment in Sovereign AI supports broader national preparedness in dual-use sectors ranging from quantum, energy, and aerospace to climate resilience and biosecurity. A narrow focus on AI as a commercial asset risks missing its wider utility as a strategic accelerator of national capability across the entire science and technology base.

6.2 Emerging Sovereignty Opportunities in Neuromorphic and Hybrid Compute

As sovereign defence institutions seek to balance strategic autonomy with long-term sustainability, a new domain of opportunity is emerging in the design and control of AI-specific hardware. While much of the current discourse on AI sovereignty focuses on datasets, models, or cloud infrastructure, the future of AI will be shaped just as profoundly by the underlying compute architectures that enable those systems to function in the field. Crucially, these architectures are no longer passive infrastructure. They are becoming active participants in the design, safety, and behaviour of AI systems. This is particularly true in the growing field of neuromorphic and hybrid computing.

It is essential to distinguish between the computational demands of training large models and those of deploying or fine-tuning them. While model training may require weeks of high-throughput GPU clusters, inference and lightweight re-alignment typically run on smaller, more portable infrastructure. Sovereign strategy should reflect this asymmetry ensuring robust training capacity where necessary, while maximising mobility and responsiveness at the edge.

Neuromorphic computing refers to a class of architectures that are inspired by the structure and function of biological neural systems. Rather than processing information sequentially or relying on energy-intensive matrix operations typical of conventional GPUs, neuromorphic chips use networks of spiking neurons to process information asynchronously and in parallel. These chips are designed for high efficiency, low latency, and adaptive learning, making them uniquely suited for event driven, time sensitive defence applications such as persistent surveillance, adaptive targeting, embedded autonomy, and multi-modal sensor fusion at the edge. A notable example of early national investment in this space is the UK's [Neuromorphic Computing Centre](#) which

focuses on brain-inspired computational architectures for energy-efficient, real-time processing. The centre, housed within Aston Institute of Photonic Technologies (AIPT), represents a convergence of neuroscience, photonics, and AI, offering a unique platform for sovereign research into spiking neural networks, embedded cognition, and adaptive signal processing. Centres such as this offer states not only technological insight but a strategic foothold in the design of mission specific AI hardware, a crucial step in achieving full-spectrum AI sovereignty.

Other emerging architectures include [hybrid analog-digital processors](#), which use analog computation to accelerate inference with greater efficiency, and photonic processors, which utilise light rather than electricity to perform high-speed parallel computation. These platforms offer domain-specific performance advantages for particular military tasks: secure satellite communications, autonomous ISR processing, or embedded signal intelligence. In each case, the performance, predictability, and controllability of the system is not a property of the software alone, but of its entanglement with bespoke hardware.

For states that wish to maintain operational and ethical authority over their AI systems, this presents both a challenge and a strategic opening. If these hardware systems are developed abroad, governed by opaque IP regimes, or produced in jurisdictions with conflicting geopolitical commitments, the resulting AI systems, no matter how carefully designed or audited, may be vulnerable to hidden dependencies, verification limits, or update constraints. On the other hand, if states act now to shape the development of these architectures, they can embed sovereign principles into the physical substrate of AI capability. Safety, override logic, inference logging, or lawful command interfaces can be implemented not only in code, but in silicon.

This requires a shift in how Sovereign AI infrastructure is conceptualised. Investment must move beyond model training clusters and general purpose data centres, and toward secure, mission-specific compute environments capable of hosting next generation architectures. Public-private partnerships with chip design firms, academic laboratories, and trusted fabrication pathways will be essential. States should also coordinate internationally to define assurance standards for AI hardware, including red-teaming protocols, traceability of fabrication origin, and hardware-in-the-loop simulation tools.

Importantly, many critical AI systems such as ISR fusion, adversarial detection, or supply chain optimisation do not require frontier-scale compute or proprietary foundation models. Sovereign capability can be built

now using existing secure compute, open-source frameworks, and classified datasets already available to defence stakeholders. Progress does not need to wait for large-scale GPU infrastructure; it begins with the integration of domain-specific models and cleared personnel into operational pipelines today.

The sovereignty of tomorrow will not be secured solely through governance frameworks or model registries. It will be secured through the ability to define, design, and control the compute architectures upon which national judgement is exercised. Those who can shape the substrate will shape the system. Those who cannot will be constrained by the assumptions, values, and strategic priorities embedded in architectures they did not design.

6.3 Model Sovereignty

Model sovereignty refers to the ability to design, govern, and adapt the architecture, training objectives, and alignment logic of AI systems deployed in defence and national security domains. In high-consequence environments such as targeting, command support, and cyber response, control at the model layer is essential to ensuring that AI behaviours remain aligned with domestic policy, lawful intent, and operational expectations.

Sovereignty in this context does not require bespoke model development in all domains. Open-source architectures or commercial models with transparent weights may suffice for certain use cases provided they can be independently audited, re-tuned, and deployed under strict technical and legal safeguards. However, assurance is rarely binary. Full evaluation of model alignment, security, and interpretability often depends on deep knowledge of the model's training data, tuning regime, and developmental history. In many cases, externally developed models may present 'black box' risks, containing latent behaviours, untraceable failure modes, or embedded assumptions

misaligned with national doctrine.

Importantly, sovereign model stewardship is no longer defined solely by architecture or parameter count. Capability is increasingly shaped by algorithmic techniques, preference learning loops, and dynamic tool use. Reinforcement Learning from Human Feedback (RLHF), Reinforcement Learning from AI Feedback (RLAIF), and fine-grained test-time optimisation allow models to outperform their pretraining baseline through alignment, reward shaping, and adaptive inference. These techniques raise sovereign control questions not just about system outputs, but about who generates the training data, who defines reward signals, and who evaluates compliance with legal or strategic constraints.

While sovereign foundation-model capabilities may be pursued at the national level as a matter of long-term strategic ambition, the operational requirements of the British Army and comparable forces are better served by smaller, domain-tuned models.

Trillion-parameter large language models (LLMs), while valuable for experimentation and centralised applications, are ill-suited to forward-deployed use due to their demands on power, latency, verification, and bandwidth. Tactical deployment environments require mid-sized, verifiable models that can operate under contested network conditions and be hosted on secure, low-footprint infrastructure.

Prioritising this scale of model development enables faster deployment cycles, reduces risk from misaligned inference, and enhances trust in AI outputs within command chains. This approach also supports alignment with the MOD Defence AI Strategy, which emphasises deployable, governed, and legally accountable systems.

6.4 Secure-by-Design Processors and Hardware Assurance

The push toward sovereign control over emerging compute architectures must be matched by a clear and sustained commitment to security at the hardware level, particularly in the era of AI co-processors, neuromorphic systems, and mission-deployed inference platforms. Sovereign AI discourse has tended to focus on data and model governance, however the foundational reality is that AI runs on processors assembled from intellectual property (IP) cores sourced from a multiplicity of vendors, many of whom are governed by commercial rather than national security priorities.

While neuromorphic, analog-digital hybrid, and edge-optimised processors offer significant advantages in terms of latency, energy efficiency, and embedded autonomy, they also introduce new challenges for assurance and oversight. One notable trade-off is a reduction in inference explainability: the internal processes of such architectures often lack the transparency and stepwise logic of more traditional AI systems. This complicates efforts to audit decisions, verify model behaviour, or reconstruct reasoning paths in post-mission review, particularly in safety-critical or legally accountable contexts. Sovereign governance in this domain must therefore extend beyond hardware provenance to include dedicated investment in interpretability tooling and simulation-based validation techniques.

The **United Kingdom** has taken an early leadership position through the [Digital Security by Design](#) (DSbD) programme, a cross-sector initiative funded by UK Research and Innovation. This programme brings together public research institutions and commercial industry to develop secure by default and by design computing systems, with a specific focus on preventing common vulnerabilities through hardware level safeguards. At the centre of this initiative is the [Morello](#) prototype platform, developed in partnership with Arm and the University of Cambridge. Morello is based on the CHERI (Capability Hardware Enhanced RISC Instructions) architecture, which introduces memory safe access controls and fine grained hardware enforced isolation mechanisms. These features are designed not merely to prevent low-level software errors, but to provide verifiable enforcement of security boundaries, making it highly relevant for AI systems used in sensitive or contested domains.

Across the **European Union**, the strategic imperative of semiconductor sovereignty has been formalised through the [European Chips Act](#). With over €43 billion of public and private funding committed, the Chips Act aims to increase the EU's share of global chip manufacturing, reduce dependency on non European supply chains, and develop a new generation of trusted, high performance processors.

Among its goals are the creation of secure edge AI chips, open-source architectures with hardware security extensions, and photonic processors capable of supporting AI-enabled automation in both civil and dual-use contexts. Complementing this, the CEPS [“Eurostack”](#) proposal highlights that Europe currently imports approximately 80% of its AI infrastructure stack—underlining the need for modular, sovereign alternatives across the digital value chain and reinforcing the principle that architectural control is a prerequisite for strategic autonomy.

Japan has launched the [RAPIDUS](#) initiative, with substantial government backing to develop a domestic 2-nanometre fabrication capability by the latter part of this decade. The objective is not simply industrial competitiveness, but national control over critical infrastructure and compute capacity, particularly in high-end AI and defence-relevant applications. RAPIDUS represents Japan’s strategic response to the geographic concentration of fabrication capacity and seeks to ensure that high-assurance chip production remains under trusted jurisdiction.

India has also moved to establish greater hardware sovereignty through its [Semiconductor Mission](#), a national programme designed to build end-to-end capability in design, verification, and fabrication of chips for defence, space, and public security. This initiative includes the development of secure embedded processors and trusted intellectual property cores, reducing dependence on foreign suppliers and enabling mission-specific chip architectures with known provenance.

Australia, while earlier in its journey, has begun aligning with [AUKUS Pillar II](#) to explore trusted hardware pathways for AI, quantum, and cyber defence technologies. There is growing interest within the Australian security community in hardware-software co-design principles that could support the deployment of AI at the tactical edge, especially in autonomous ISR platforms and maritime sensing environments. Together, these programmes illustrate a global trend.

The future of Sovereign AI will not rest solely on who trains the models or governs the data, but on who defines and secures the chips that power critical systems. This is not simply a matter of technological preference but of operational resilience, legal accountability, and strategic freedom of action. Secure-by-design computing is therefore not a narrow engineering challenge, it is a geopolitical one. In military and intelligence environments, the risks of hardware compromise are non trivial. AI-enabled systems used in theatre, such as tactical ISR drones, edge inference devices, or targeting processors are often assembled from componentised hardware IP sourced globally, with limited visibility into firmware, logic behaviour, or update provenance. Even in highly classified environments, the defence sector remains reliant on commercial semiconductor supply chains, which are opaque, fragmented, and increasingly geopolitical.

Moreover, uptake of hardware level security enhancements has historically been slow, particularly in commercial chip design, due to performance and cost constraints. Protective features such as capability based memory protection, logic confinement, or secure boot pathways take up valuable silicon real estate and reduce margins. The result is a persistent security versus efficiency trade-off, one that commercial providers are not structurally incentivised to resolve without state driven procurement standards or R&D partnerships. To address this, sovereign states must treat chip level assurance not as a secondary procurement criterion but as a primary strategic requirement. Secure-by-design architectures should be mandated for AI processors in high-risk domains, and efforts such as DSbD should be expanded to include mission-specific AI hardware, with threat models that reflect the unique risks posed by inference compromise, embedded manipulation, or sensor level corruption. Red-teaming, firmware traceability, and hardware provenance validation should be part of national assurance pipelines, not post deployment patches.

Defence institutions should also consider establishing trusted silicon enclaves within their supply chains, partnering with fabrication partners under reciprocal security agreements and independent verification regimes. These measures are essential not only for protecting AI performance but for preserving lawful command and operational independence at the hardware level.

Sovereign AI requires sovereign trust in the chips on which it runs. Where compute is compromised, no level of model auditability or data governance can restore system integrity. Security must therefore begin not with software updates or application-layer oversight, but with the substrate, the physical and logical foundations of computational authority.

6.5 Beyond the Capital Cost

Sovereign AI must be sustained through ongoing operational expenditure. This includes the costs of power, physical and cyber security, personnel, system maintenance, red-teaming, retraining cycles, and the integration of models with live operational systems. Maintaining such a capability also involves continuous oversight of model drift, inference governance, and mission-specific adaptation.

Operational staffing includes sovereign software engineers, security-cleared model trainers, and dedicated oversight personnel within Defence Digital, DSTL, and forward-deployed commands. These costs, though substantial, are predictable, scalable, and yield long-term value across multiple Defence applications, effectively amortising investment across strategic capabilities. Compared with the long-term risks and expenses associated with outsourcing critical AI capabilities to external vendors, particularly those operating under opaque licensing regimes or foreign legal frameworks, a sovereign approach offers superior security, transparency, and long-term cost stability.

The question of feasibility, however, is not only one of cost. It is also one of industrial capacity. The UK does not currently possess end-to-end Sovereign AI capability across the full technology stack. It does not manufacture high end AI chips. It does not host hyperscale cloud environments capable of serving MOD-wide AI operations from sovereign soil. It relies on foreign firms for cloud compute, hosting, and in some cases, advanced model development. These are

vulnerabilities. But they are not disqualifications. In fact, the UK possesses unique industrial advantages in precisely those parts of the AI stack most relevant to sovereign defence application.

The UK's academic AI community anchored by the Alan Turing Institute, the universities of Oxford, Cambridge, Edinburgh, Imperial, and UCL is globally recognised for excellence in algorithmic design, machine learning assurance, and statistical robustness.

The UK benefits from a growing ecosystem of SMEs capable of contributing to Sovereign AI development across defence and national security contexts. The Defence AI Centre ([DAIC](#)) has helped surface this capability through initiatives such as the AI Expert Group, which convenes technical leaders and innovators across industry, academia, and government. Additionally, the DAIC Connect programme, run in partnership with [Chief Disruptor](#), has drawn participation from a wide range of AI SMEs with relevant capabilities in modelling, assurance, interface design, data curation, and trusted autonomy.

On 25 February 2025, DAIC Connect in partnership with Chief Disruptor, hosted the second event in the City of London. The event brought together stakeholders from across government, industry, and academia to advance the UK's defence AI ecosystem. Its purpose was to deepen understanding of the Defence AI market, share strategic updates from the Ministry of Defence, and foster practical collaboration between AI innovators and national security institutions.



The event featured keynote addresses by senior MOD officials, updates on ongoing initiatives, and panel discussions focused on the operational challenges and strategic opportunities associated with AI integration in defence. Breakout sessions enabled in-depth dialogue on technical barriers, assurance, and deployment at scale, while dedicated networking segments helped connect capability providers with programme leads.

A diverse range of UK-based SMEs and technology firms participated, reflecting the growing breadth of the Sovereign AI industrial base. These companies brought forward capabilities spanning AI safety, trusted autonomy, real-time analytics, and

defence-grade machine learning tooling, demonstrating the strategic depth of the UK's domestic innovation landscape.

If supported through appropriate funding, validation, and risk-calibrated adoption processes, this community could underpin a UK-aligned AI industrial base that is both sovereign and exportable across trusted alliances. Collectively, these SMEs and defence primes contribute to the UK's strategic objective of achieving Sovereign AI capabilities. Their efforts ensure that the UK maintains control over critical technologies, reduces reliance on foreign entities, and enhances the resilience of national defence systems.

AI Assurance, Modelling & Safety

Advai

Advai specialises in AI robustness, adversarial testing, and assurance, helping defence and critical systems identify vulnerabilities in machine learning pipelines.

W&B

Weights & Biases provides a widely adopted suite of tools for tracking, visualising, and reproducing AI experiments. Supports rigorous model development, testing, and collaboration.



Mind Foundry, an Oxford University spinout, delivering mission-critical AI systems with a strong focus on responsible deployment, interpretability, and performance in complex environments.



Literal Labs develops symbolic and logic-based AI systems designed to be transparent, verifiable, and energy-efficient providing alternatives to opaque deep learning approaches.

Defence AI Platforms, Autonomy & Mission Systems

Helsing

Helsing is a defence software company providing AI-enabled systems that process sensor data, support decision-making, and integrate with military hardware in real time.

OXFORD DYNAMICS

Oxford Dynamics focuses on signal intelligence and AI-based behavioural modelling. Supports the development of secure, adaptive defence technologies.



Shield Reply (Reply Group) delivers secure AI and large language model (LLM) integration tailored for defence, with a focus on sovereign governance, deployment security, and trusted interfaces.

Vizgard

Vizgard specialises in AI-enabled surveillance, visual threat detection, and counter-drone solutions. Builds edge-AI platforms for situational awareness and autonomy.



Delian provides decision intelligence platforms for defence operations, with expertise in AI-driven situational awareness, data fusion, and autonomous control.



Archangel Autonomy develops autonomous edge AI systems for long-endurance, low-power surveillance in disconnected environments. Deployed for logistics, border security, and tactical ISR.

Real-Time, Edge & Infrastructure AI

Zenith Vector is an emerging player in secure AI infrastructure, offering tailored tools for deploying scalable, mission-specific AI across edge environments.

VANTIQ

Vantig enables rapid development of real-time, event-driven AI applications for high-tempo operational contexts, including logistics and asset tracking.



InfraStar delivers infrastructure-grade AI solutions for defence and critical systems. Focuses on resilience, deployment control, and cross-domain orchestration.



Periphery is an easy to embed military-grade threat management system for IoT manufacturers.

AI Integrators, Decision Support & Trusted Intelligence



Intellium AI Offers trusted AI development and consultancy services across sectors. Delivers end-to-end AI pipelines with a focus on explainability, safety, and deployment control.



TRM Labs provides blockchain intelligence tools for detecting financial crime and threat finance. Used in national security, sanctions enforcement, and crypto investigations.



Great Wave AI works on trusted intelligence and AI-enabled decision support, with applications in government and regulated industries.



Quantexa specialises in contextual decision intelligence, using graph AI to map relationships, identify risks, and support threat analysis across domains.



Cineon builds emotion AI systems for training, simulation, and human performance monitoring. Applies human-centred modelling to security, defence, and emergency services.

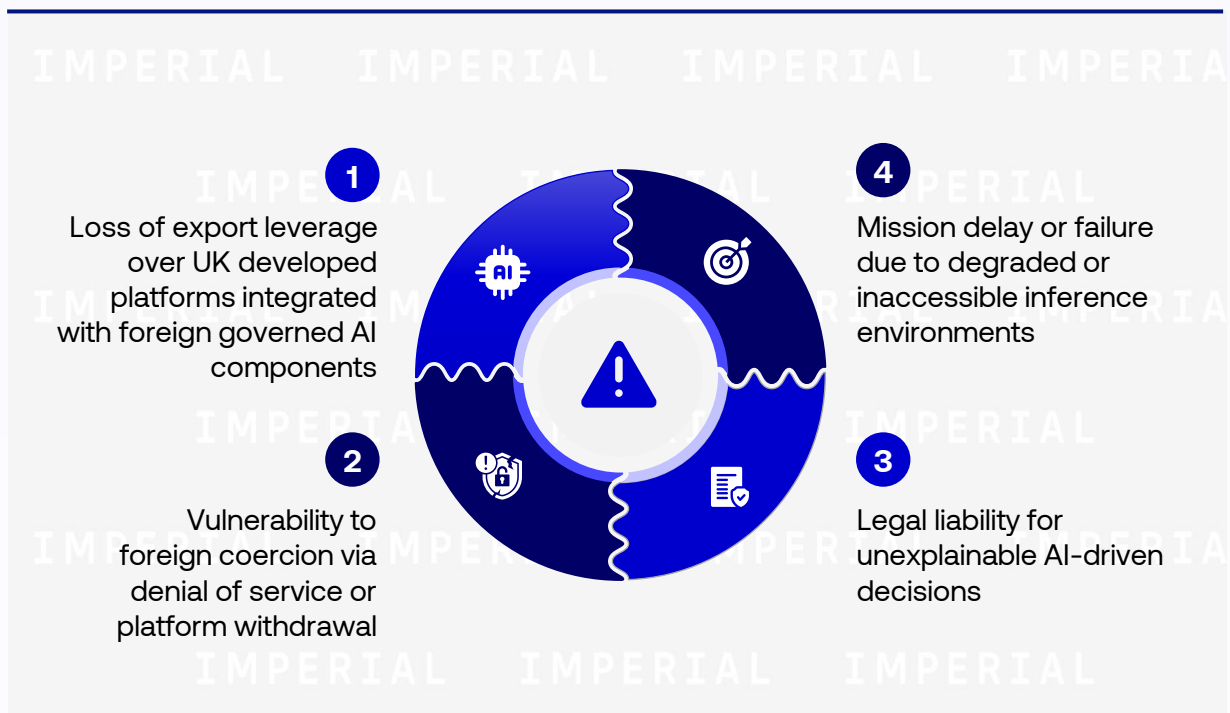
6.6 Constraints and dependencies

The primary constraint lies in hardware and hyperscale infrastructure. The UK currently depends on NVIDIA and AMD for advanced GPU supply, and on Taiwan and South Korea for chip fabrication (TSMC and Samsung). Hyperscale compute is dominated by US-headquartered cloud vendors, including AWS and Microsoft Azure, who operate under US jurisdictional frameworks. This dependency introduces both strategic exposure and legal risk. For instance, cloud environments hosted by foreign providers may be subject to lawful access orders, commercial prioritisation shifts, or data residency conflicts that prevent MOD from maintaining continuous or exclusive control over defence-critical models.

The way forward is not to eliminate these dependencies, which would be prohibitively expensive and industrially unrealistic. Instead,

the UK must mitigate them through diversification, modularity, and intelligent public-private partnerships. Stockpiling and procurement diversification of AI accelerator hardware, co-investment in European chip initiatives, and the creation of MOD-managed sovereign enclaves within Crown Hosting environments are all within reach. Additionally, sovereign deployment patterns such as edge inference, containerised retraining, and hybrid federated learning architectures offer the MOD the ability to exercise functional sovereignty even when relying on foreign-designed silicon or shared supply chains.

Ultimately, Sovereign AI should be understood as a strategic asset with cost avoidance benefits. The absence of sovereign control in key domains creates downstream risks such as:



These costs are difficult to price but their strategic gravity is unquestionable. By contrast, the cost of building targeted Sovereign AI capability in key mission domains is both known and bounded. It is, in

short, a form of strategic insurance, paid not to protect against failure, but to preserve the nation's ability to decide, act, and lead under its own authority.

6.7 Strategic Enablers of Sovereign AI – Talent, Clearance, and Workforce Development

Sovereign AI cannot be credibly pursued without sovereign talent. While infrastructure, legal frameworks, and governance models define the architecture of control, it is cleared, capable personnel who operationalise sovereignty. **Strategic autonomy in the digital age depends not only on who builds systems, but on who is trusted to access, adapt, and govern them under classified and contested conditions.** This reliance on skilled and security-vetted personnel presents a structural challenge for many governments. In particular, the defence AI sector suffers from critical gaps in available talent cleared to operate within sensitive environments. Startups and SMEs where much frontier innovation originates are often excluded from sovereign contracts due to slow or opaque clearance processes.

This disconnect undermines both innovation and resilience.

To address this, **states must treat security-cleared AI talent as a national capability in its own right.** A promising model exists in the UK's [NCSC i100](#) initiative, which embeds pre-cleared private sector experts into sensitive cyber defence missions. The i100 offers a working example of how governments can integrate non-traditional and agile talent into mission-relevant roles without compromising assurance. Participants are seconded from industry with vetted access, enabling the state to benefit from specialist insight while preserving institutional control.

Governments should build on this logic by establishing a Sovereign AI Talent Reserve: a pre-cleared, multidisciplinary pool of AI engineers, model auditors, assurance specialists, and legal advisors who can be deployed flexibly across sovereign programmes. Such a pool could be governed under a “Sovereign AI League” model, enabling trusted engagement, community exchange, and time-boxed contribution mechanisms for cleared personnel.

At the same time, Sovereign AI demands a broader pipeline of career-ready talent. The UK’s Defence AI Strategy (2022) calls for Defence to become “AI ready” by investing in upskilling, recruiting, and developing specialist roles. The British Army’s 2023 [Approach to Artificial Intelligence](#) echoes this, emphasising baseline digital literacy and advanced capability tracks stating that “*The Army will be AI ready when relevant parts of the workforce are enabled with a baseline AI digital literacy, data quality is enhanced,*

access to technology and established relevant processes required to deliver assured, safe, and responsible AI”.

To meet these objectives, governments must expand AI-dedicated career streams across Defence Digital, DSTL, and operational commands, building paths for data scientists, assurance engineers, human-machine teaming specialists, and digital operations planners.

These roles must be underpinned by a continuous reskilling ecosystem, integrating military education programmes, academic partnerships, and industrial placements. They must also be future-proofed through reservist pathways, modular training pipelines, and the integration of AI roles into established force design models. Without these enablers, AI sovereignty will remain aspirational.

To institutionalise this capability, governments should:

- ✓ Commission a review of clearance barriers for AI-specific roles, particularly for SMEs and international contributors
- ✓ Develop a pre-clearance talent pipeline, with staged vetting, provisional access, and agile deployment structures modelled on the i100 programme
- ✓ Align this pool with assurance, audit, and oversight functions central to Sovereign AI deployment
- ✓ Expand formal AI career streams within defence institutions, backed by ongoing training and joint academic-industry programmes
- ✓ Explore cross-border eligibility options for diaspora and allied experts operating under national governance frameworks

AI sovereignty is ultimately delivered by people. Without the ability to train, clear, and retain skilled personnel who can govern

systems in real time, no institutional structure or policy posture can remain effective under pressure.

6.8 Data Advantage in Defence: From Dark Data to Sweet Spot Models

While the UK may not be a leader in scaling general purpose frontier models, it possesses a distinct and underexploited advantage in a different class of data: mission-specific, high-fidelity, and operationally sensitive information. Often referred to as dark data, these datasets are held within secure defence repositories but remain underused in AI development due to their classification, modality complexity, or the absence of dedicated model pipelines.

This category includes sonar and acoustic data from undersea platforms, electromagnetic and RF emissions from electronic warfare systems, battle telemetry and control system logs, high-grade sensor fusion streams from ISR platforms, and simulated conflict scenarios generated through defence wargaming environments. These are not only rich sources of structured information, but they are often unavailable or unusable to commercial actors due to legal, ethical, and national security constraints.

These datasets represent what can be called the “sweet spot” for Sovereign AI: domains where scale is less important than control, relevance, and precision. Models trained on defence specific datasets can be smaller, more targeted, and more tightly aligned with operational doctrine. These models offer a route to sovereign capability that does not require competing head-to-head with commercial AI labs, but instead leverages the UK’s strategic position as a generator and custodian of unique, high-trust data.

This opens a viable and defensible path to sovereign advantage. By focusing on bespoke, classified, or semi-structured data that is already under MOD governance, the UK can accelerate AI development in areas where commercial providers cannot operate, while simultaneously retaining full lifecycle control over training, alignment, deployment, and auditability. It also allows AI systems to be tuned to UK-specific mission requirements, legal frameworks, and force

integration standards, thereby improving both performance and assurance.

Realising this opportunity will require deliberate investment in data curation, access protocols, and secure compute environments. It will also demand that defence institutions treat data not merely as a by-product of operations, but as a sovereign asset, an enabler of trusted autonomy, real-time decision support, and operational edge. With proper governance, these datasets can be mobilised to train AI systems that provide asymmetric advantage in high sensitivity domains where explainability, agility, and institutional trust matter more than brute scale.

To fully capitalise on the UK’s access to defence-specific datasets, the MOD and allied institutions should prioritise the fusion of UK-held classified data with real-world combat datasets from trusted partners. The war in Ukraine has produced a wealth of operationally rich telemetry such as drone ISR footage, counter-UAS logs, battlefield damage imagery, and real-time electronic warfare (EW) patterns that, if securely accessed and harmonised, could significantly enhance the training of sovereign, domain-specific AI models. These datasets represent practical, high-fidelity complements to UK sensor logs and mission data. Formalising data-sharing agreements with Ukraine and close allies would turn warfighting lessons into a tangible sovereign capability advantage. This direction is aligned with priorities outlined in the [Defence Command Paper Refresh \(2023\)](#), which stresses the value of “combat-experienced” data in accelerating AI capability delivery and reducing synthetic–real domain gaps.

Ultimately, the UK’s most promising opportunity for Sovereign AI leadership lies in exploiting the underleveraged specificity of its own dark data. The systems developed in these “sweet spot” domains will not only be sovereign by design, they will be strategically irreplicable.

6.9 Strategic Integration, UK MOD: Insights from the Sovereign AI Initiative

The UK Ministry of Defence's pursuit of targeted sovereign Artificial Intelligence should be guided not only by operational imperatives but also by coherent alignment with national AI policy principles. As we have discussed in earlier chapters, 2025 white paper Considerations Regarding Sovereign AI and National AI Policy ([Trusted AI Alliance](#)) offers a conceptual and strategic foundation that complements the defence-specific

approach outlined in this report. Integrating these principles from the Sovereign AI white paper into its planning, the MOD can ensure that its defence AI posture is not only mission-credible but also strategically harmonised with wider national interests. This alignment will support long-term resilience, cross-sector interoperability, and international leadership in trusted, sovereign defence AI.

Principle 1	Principle 2	Principle 3	Principle 4	Principle 5
Coherence Between National and Defence Sovereignty	AI Infrastructure as a National Security Asset	Accountable Sovereignty Through Legal and Ethical Auditability	Risk-Based Sovereignty Allocation	Industrial and Skills Strategy as a Sovereignty Enabler

Principle 1: Coherence Between National and Defence Sovereignty

The Sovereign AI Initiative contends that digital sovereignty must be conceived as a whole-of-government and whole-of-society endeavour. For MOD, this means Sovereign AI policy should not be stove-piped from broader national digital policy. Strategic alignment with the Cabinet Office, Department for Science, Innovation and Technology (DSIT), and the Office for AI is essential. Defence-specific capabilities such

as secure model hosting, red-teamed AI assurance frameworks, and UK-governed inference engines must interoperate with civilian initiatives around AI safety, digital infrastructure, and trust frameworks. This approach strengthens legal harmonisation, investment synergy, and unified standards across civil and military sectors. Defence AI must be a pillar of UK digital sovereignty, not an outlier.

Principle 2: AI Infrastructure as a National Security Asset

AI infrastructure, including compute clusters, sovereign data pipelines, and inference environments must be treated as foundational strategic infrastructure, akin to energy or telecommunications. The Trusted AI Alliance report suggests that future national resilience will depend on domestically governed, secure, and modular compute infrastructure. For the MOD, this principle justifies the establishment of Crown

owned AI environments that allow for high assurance development and deployment. It also underlines the urgency of diversifying hardware supply chains and contributing to UK or European chip strategy consortia. MOD's infrastructure must be survivable, scalable, and sovereign, forming a digital fortress from which mission critical AI can be trained, validated, and operated under UK jurisdiction.

Principle 3: Accountable Sovereignty Through Legal and Ethical Auditability

The foundation of democratic military power is accountability, not only in intent but in traceable, institutionalised governance. Sovereign AI within the MOD must be designed from the outset to enable full legal and ethical auditability, particularly in domains that carry kinetic consequence, such as targeting, intelligence fusion, and command and control (C2). In this context, sovereignty does not merely mean national control over infrastructure or models; it also requires that every AI-driven output can be understood, justified, and legally defended.

The Sovereign AI white paper and MOD doctrine converge on a key point: **the delegation of decision support to AI systems must not compromise the chain of accountability.** Human operators, commanders, and ministers remain legally responsible for the outcomes of defence operations. Therefore, AI systems must be structured to support this responsibility, not obscure it. This means embedding mechanisms for inference logging, source data verification, and decision traceability across the AI lifecycle.

To achieve this, **MOD must develop and institutionalise a robust Defence AI Assurance Framework.** This framework should define standards for auditability, model interpretability, and inferential oversight. It must include mandatory red teaming protocols, legal pre-authorisation pathways, and dynamic rules of engagement calibration for AI models deployed in operational theatres.

Two institutional actors are critical to delivering this framework: the MOD Legal Directorate and the Joint Doctrine Centres.

The MOD Legal Directorate provides authoritative legal guidance across operational and strategic contexts. It ensures that UK defence systems comply with the Law of Armed Conflict (LOAC), international humanitarian law, and domestic legal obligations. For AI systems, this Directorate must have oversight of model design and inference governance to certify that decisions involving force application can withstand legal scrutiny, both in domestic courts and under international law.

The Joint Doctrine Centres, under Strategic Command, are responsible for codifying how the UK Armed Forces fight. Their role is to ensure doctrinal coherence across services and domains. Integrating Sovereign AI into UK military doctrine requires that these Centres formalise the conditions under which AI may support or shape battlefield decisions, including stipulations for human-in-the-loop oversight, override authority, and operational limits on autonomous functions. Together, these institutions must collaborate to embed Sovereign AI not just in code, but in command culture. Legal accountability and doctrinal legitimacy must be hardcoded into AI-enabled systems as core design principles, not post-deployment add-ons. Only through this fusion of legal oversight and doctrinal clarity can the MOD ensure that its Sovereign AI capabilities uphold the standards of a democratic, law-bound military power.

Principle 4: Risk-Based Sovereignty Allocation

Sovereignty in AI is not a one-size-fits-all condition. It must be scaled intelligently based on mission relevance, legal sensitivity, and operational risk. This principle, championed in the Sovereign AI white paper, underpins the MOD's posture of modular,

mission-driven sovereignty. **It calls for the UK to allocate sovereign control where it matters most, rather than overextending resources on total control of all AI systems.**

To operationalise a risk-based approach, the MOD should implement a formal AI classification matrix. This tool would score AI systems across sovereignty dimensions; data control, model governance, inference location, update autonomy, and legal

traceability, enabling strategic triage. Such a framework supports informed resourcing, coherent procurement, and targeted assurance. Sovereignty becomes a calibrated posture, adapted to strategic and operational realities.

Principle 5: Industrial and Skills Strategy as a Sovereignty Enabler

True AI sovereignty cannot be purchased off the shelf. It must be built and sustained by a domestic industrial and skills base capable of supporting sovereign development, deployment, and assurance. As the Sovereign AI Initiative stresses, the foundation of sovereignty is capability: the people, organisations, and infrastructure that allow the UK to govern its own AI future.

For the MOD, this requires a deliberate defence industrial strategy aligned to Sovereign AI priorities. Strategic partnerships must be fostered with UK-based SMEs, universities, and research institutions focused on secure, verifiable, and mission-specific AI. Defence procurement pathways should prioritise dual-use innovation, modular architecture, and sovereign reusability. Export frameworks should be developed to enable UK origin AI capabilities to scale commercially without compromising national control. Critically, MOD must invest in talent. A Sovereign AI capability requires engineers, data scientists, red-teamers, and operational integrators with the clearances

and expertise to build, validate, and deploy sensitive systems. This includes battlefield inference specialists, AI operations officers, and legal technological hybrid roles that can interpret doctrine through the lens of code. A Sovereign AI career track should be created within Defence Digital and DSTL, complemented by reservist pathways and academic secondments. These roles will not be filled by passive recruitment. **New pathways must be established to identify, clear, and retain talent from academia and industry with high-trust, deployable profiles.**

Sovereign AI must become a central pillar of the Defence and Security Industrial Strategy (DSIS) and the wider national AI Skills Strategy. Without this human and industrial foundation, the UK risks becoming a passive consumer of AI systems shaped by others' values, assumptions, and interests. With it, the UK can lead not only in ethical military AI but in building a resilient, sovereign digital defence economy.

6.10 Safety, Theory of Control and National Assurability

In a strategic context, safety is sovereignty operationalised, it enables a state to retain lawful, accountable, and effective command over AI-enabled capabilities, even in the most adversarial or uncertain environments.

Without the capacity to explain, test, and control what AI systems do, whether they support ISR, command decision-making, cyber defence, or tactical autonomy, states forfeit strategic agency, operational assurance, and legal credibility. Yet across

many jurisdictions, the institutional capacity to assure AI safety remains underdeveloped. Technical due diligence is often confined to procurement audits, while red-teaming, interpretability, and system level verifications are inconsistently applied. There is a notable asymmetry between the speed at which AI capabilities are being deployed in defence contexts and the maturity of the tools, theories, and personnel required to ensure their safe operation.

This is a structural vulnerability. Defence organisations increasingly rely on ensembles of interacting AI systems, from drones and targeting platforms to cyber sensors and C2 networks. These systems must interoperate, adapt, and respond in real-time under contested conditions. Even where individual components are validated, their joint behaviour can be unstable, unpredictable, or legally problematic. Without the ability to model, simulate, and verify these dynamics, states risk deploying systems that undermine rather than strengthen operational integrity.

Safety in this context must be treated as a sovereign capability, one that is developed, institutionalised, and retained within the national security ecosystem. It requires investment in both applied tools and foundational theory. A credible national approach to AI safety must be grounded in technical disciplines capable of rendering system behaviour intelligible, controllable, and resilient. Among these, several areas demand immediate investment and institutional focus.

First is the field of mechanistic interpretability, which seeks to uncover and explain the internal reasoning processes of AI systems. It is not sufficient to observe the outputs of a model, sovereign assurance requires insight into how those outputs were produced, what internal representations or pathways led to a given recommendation or decision, and how those might shift under new or adversarial inputs. This capacity is vital in mission critical systems, where inference errors must be traceable and explainable under legal or operational review. Second is the use of formal verification methods, tools that apply mathematical, symbolic, or logic based reasoning to prove that certain behaviours or safety properties hold across all valid model states or input conditions. These techniques allow system developers and commanders alike to assert, with rigour, that a system will not exceed defined operational parameters, violate

engagement rules, or produce contradictory outputs under specific conditions.

A third area of critical importance is adversarial robustness. AI systems must be resilient to intentional manipulation through adversarial inputs such as spoofed imagery, corrupted data, prompt injection, or inference attacks. Without hardened defences, systems may misclassify threats, misattribute actions, or propagate tainted information through decision pipelines, with potentially catastrophic consequences.

Equally essential is the development of ensemble assurance frameworks, which address the growing complexity of AI systems operating in coordination. As defence institutions deploy AI across ISR, targeting, logistics, and C2 platforms, the need to understand the emergent behaviours of these systems in concert becomes paramount. What is predictable in isolation may be unstable in interaction. Sovereign assurance must extend not only to components, but to the dynamics of system-of-systems behaviour.

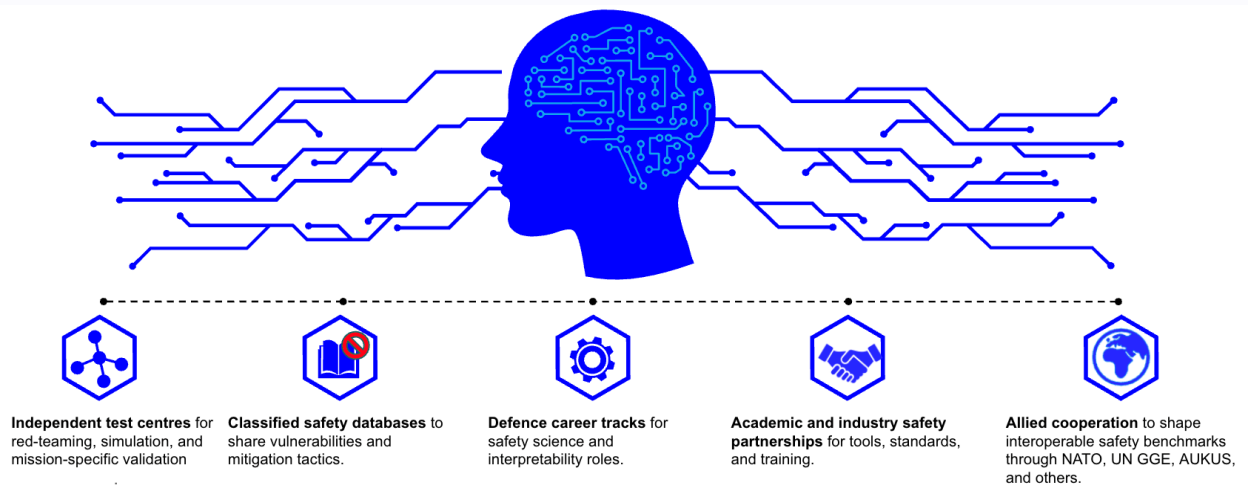
Finally, safe deployment requires the design and implementation of fail-safe architectures and override mechanisms. These are not generic kill-switches, but tailored affordances that allow authorised human actors or supervisory systems to interrupt, suspend, or reconfigure AI behaviours in response to malfunction, adversarial manipulation, or unexpected environmental change. Such mechanisms, whether through rollback protocols, mission-specific constraints, or dynamic parameter resets, are the ultimate safeguard of lawful and accountable military command.

Together, these disciplines form the technical foundation of AI sovereignty. They enable states to move beyond trust in systems to trust in their understanding of those systems, a necessary condition for responsible deployment and credible command.

Such capabilities cannot be outsourced. Safety is not simply a matter of contract compliance; it is an inherently sovereign function, just as doctrine development, intelligence vetting, and rules of engagement are sovereign functions. States that delegate

safety to unverified commercial providers or rely entirely on opaque black-box systems risk losing not only command but legitimacy, unable to justify actions taken under the influence of systems they do not fully understand.

The Trusted AI Alliance proposes the implementation of national assurance frameworks that treat AI safety as a first order defence priority. These frameworks should include:



Sovereign AI systems must not only be under national jurisdiction, they must be under national comprehension. A model that cannot be explained cannot be audited. A system that cannot be assured cannot be governed. And an ensemble that cannot be controlled cannot be trusted, not by

commanders, legislators, or the public. Safety must be integrated not only into model design and deployment, but into strategic doctrine. Defence ministries must be able to answer, with confidence and evidence, fundamental questions:

1

What will this system do under stress?

2

Can we verify its behaviour?

3

Can we safely interrupt or override it?

These are not technical luxuries; they are operational and ethical necessities.

Ultimately, sovereignty in AI-enabled defence does not reside in code ownership or infrastructure control alone. It resides in assurability, the ability to predict and constrain what systems will do, to justify their

outputs under legal scrutiny, and to retain meaningful human command over decisions that carry life-and-death consequences. Without this, national authority is hollow. With it, states retain the most fundamental attribute of sovereign power: the ability to decide and to be accountable for what is done in their name.

7. Strategic Synthesis and Posture Recommendation

As demonstrated throughout this white paper, the Ministry of Defence cannot fulfil its legal obligations, maintain operational credibility, or uphold alliance trust unless it asserts targeted sovereign control over the AI systems that inform or execute military action. The strategic rationale for Sovereign AI established in Chapter 4 rests on five foundational imperatives: operational assurance, legal and ethical accountability, strategic autonomy, coalition credibility, and domestic industrial resilience. Each of these imperatives intersects with a single organising principle: the UK must be able to understand, direct, and take responsibility for the actions of its AI-enabled systems, particularly in domains where the consequences of error, compromise, or dependency are irreversibly high.

The question is not whether Sovereign AI is necessary, it is how it can be structured, prioritised, and institutionalised in a way that is both strategically credible and economically feasible.

The answer, as developed across this paper, lies in a posture of **mission-driven, modular sovereignty**: a differentiated model of control that aligns the degree of sovereign oversight with the operational risk, legal exposure, and strategic significance of each AI application. This posture avoids both extremes. It rejects total autarky, which is fiscally and industrially unsustainable, and it avoids blind dependency on foreign

platforms, which introduces unacceptable liabilities in crisis scenarios. Instead, it proposes a sovereignty gradient calibrated to function, consequence, and context.

This posture is operationalised through the six interdependent governance dimensions introduced in Section 2.1 Data Governance, Model Governance, Training and Alignment Governance, Compute Governance, Operational Governance, and Legal and Ethical Governance. These dimensions provide a structured and actionable framework for determining where sovereign control must be assertive and where it may be permissive. They move the concept of sovereignty beyond ideological assertion and ground it in the institutional mechanisms by which control is maintained, accountability is discharged, and resilience is preserved.

This posture is operationalised through the six interdependent governance dimensions introduced in Section 2.1 Data Governance, Model Governance, Training and Alignment Governance, Compute Governance, Operational Governance, and Legal and Ethical Governance. These dimensions provide a structured and actionable framework for determining where sovereign control must be assertive and where it may be permissive. They move the concept of sovereignty beyond ideological assertion and ground it in the institutional mechanisms by which control is maintained, accountability is discharged, and resilience is preserved.

The imperative to apply these dimensions comprehensively is clearest in the high-risk operational domains analysed in Chapter 5:

ISR Fusion

In ISR Fusion, AI systems must be able to process and fuse multi-source intelligence to support lawful targeting. Here, legal and ethical governance is paramount, and sovereignty over model inference behaviour and data provenance is non-negotiable.

Command Decision Support

In Command Decision Support, AI tools shape operational and strategic judgement. Misaligned systems could bias decisions or obscure responsibility. Sovereign control over training alignment and model behaviour is essential to maintain doctrinal coherence and ensure decision accountability.

Defensive Cyber Operations

In Defensive Cyber Operations, AI systems operate as first responders to hostile incursions. Without sovereign update and operational governance, these systems may be compromised or delayed in adapting to threats. Compute sovereignty and red-teamed model assurance are central.

Tactical Autonomy and Embedded Inference

In Tactical Autonomy and Embedded Inference, AI systems enable autonomous platforms to operate under contested conditions. Sovereign control over embedded inference logic, fail-safe mechanisms, and ethical constraints ensures that autonomy does not become irresponsibility.

Cognitive Security and Information Operations

In Cognitive Security and Information Operations, AI is used to monitor, interpret, and counter adversarial influence. Sovereignty here is not only technical, it is normative. Systems must operate under UK definitions of manipulation and democratic risk, not foreign moderation policies.

Each of these domains involves functions that touch directly on the application of force, the attribution of intent, or the preservation of democratic legitimacy. In such contexts, AI systems must not merely be performant, they must be auditable, governable, and responsive to UK authority across their entire lifecycle. Sovereignty is not simply a matter of who builds the model, but who can adapt it in the moment of operational need, who can justify its outputs under legal scrutiny, and who can decommission or override it when strategic, ethical, or political circumstances demand.

By contrast, in lower-risk domains such as HR analytics, logistics optimisation (peacetime), or enterprise management the strategic consequences of failure or compromise are more contained. As explained in earlier chapters, these domains may leverage commercial AI solutions, provided that sovereignty is maintained through contractual governance, data minimisation, and fallback mechanisms. Sovereign oversight in these contexts is procedural rather than developmental. This preserves agility and economic efficiency while safeguarding strategic flexibility.

This differentiated posture delivers key strategic benefits such as



Defence institutions must move beyond ad hoc development and establish coordinated frameworks for Sovereign AI governance. In the UK context, this may include a dedicated Directorate model, as proposed herein, to unify assurance, model registries, and legal oversight.

This paper recommends the creation of a Defence AI Sovereignty Directorate, reporting jointly to Strategic Command and the Chief Scientific Adviser, with formal integration with the Defence AI Centre, Crown Hosting Data Services, and the MOD Legal Directorate.

This Directorate should be tasked to

-  Define sovereignty benchmarks and risk thresholds across AI systems and procurement categories
-  Maintain a classified registry of sovereign models and AI components used in high-risk applications
-  Certify systems against sovereign assurance standards, including legal auditability and mission-aligned training
-  Serve as the UK's focal point for alliance-level collaboration on Sovereign AI interoperability, assurance, and joint certification
-  Lead MOD's AI red-teaming, validation, and alignment assessments across operational theatres, building on existing capabilities, such as those within DSTL, while preserving and integrating the specialist personnel already driving these functions forward

This organisational structure must be matched by policy integration.

- 01 Sovereignty metrics must be embedded into the Defence Equipment Plan, Defence Digital Strategy, and all DSIS evaluations
- 02 Sovereign AI capability must become a standard of force readiness just as much as physical deployability, C2 resilience, or cyber hardening
- 03 More fundamentally, Sovereign AI must be understood not as a procurement choice, but as a strategic function of national defence doctrine.

Without such a posture, the UK risks becoming a consumer of strategic cognition, dependent on external actors for the systems that interpret threats, structure decisions, and even direct action in moments of national consequence. In such a condition, military capability may be preserved, but military authority is diminished.

With Sovereign AI, the UK affirms its position as a law-bound, accountable, and strategically autonomous power. It maintains not just the ability to act, but the sovereign responsibility to decide, under its own terms, through its own systems, and in accordance with its own values. This is not simply a matter of technical design. It is a matter of national command.

7.1 Operationalising Sovereignty: The Role of Metrics

For a posture of modular, mission-driven sovereignty to function as more than strategic intent, it must be translated into an operational framework through which AI systems can be evaluated, governed, and assured. This translation requires the development of sovereignty metrics, structured, repeatable criteria by which institutions can assess the degree of sovereign control exercised over a given AI capability. Such metrics form the analytical substrate upon which capability classification, risk triage, procurement oversight, and assurance pathways are built. The proposed Defence AI Sovereignty Directorate would be tasked not only with setting policy direction and institutional accountability, but with defining and maintaining the technical and organisational metrics through which sovereignty is expressed and enforced across the AI

lifecycle. These metrics should not be narrowly technical. They must encompass the full arc of system development, deployment, and governance, from data and model design to infrastructure, legal accountability, and human integration. Sovereignty in AI is inherently multidimensional. No single variable can determine whether a system is sovereign. Rather, sovereignty must be understood as a gradient, composed of overlapping forms of control, assurance, and national responsibility. **The metrics must therefore be calibrated across six foundational dimensions, each corresponding to the sovereignty dimensions outlined in Chapter 2:** data governance, model governance, training and alignment governance, compute governance, operational governance, and legal and ethical governance.

Within each of these dimensions, specific indicators can be defined to assess the level of sovereign control. For instance, in the domain of data governance, relevant indicators might include the proportion of training and inference data sourced under national authority, the robustness of data provenance assurance mechanisms, and the extent to which data curation processes are auditable and mission specific. A system that relies heavily on foreign labelled datasets lacks traceability, or cannot validate the legality of its training data would score poorly on this axis, regardless of its performance characteristics.

Similarly, in model governance, sovereignty may be assessed through the degree of access and control over model architecture, weights, and behavioural tuning. Metrics would consider whether the model can be independently audited or explained, whether updates and tuning are conducted under sovereign policy constraints, and whether mission-specific performance can be verified by national technical authorities. Systems built with closed-source models, governed by external platforms, or lacking explainable decision logic would pose risks of uncontrollable behaviour and external influence.

In the area of training and alignment governance, indicators might include whether the alignment objectives and reward functions were set in accordance with national legal frameworks, whether the system can be realigned without external dependency, and whether training regimes reflect operational realities rather than generic benchmark optimisation. Sovereign AI systems must be aligned not only to performance metrics but to strategic norms, legal doctrine, and command intent.

Compute governance, perhaps the most structurally overlooked dimension, evaluates where and how systems are hosted and executed. Metrics here include the extent of control over physical infrastructure, supply

chain transparency in hardware components, security certification of compute environments, and the ability to deploy models under secure conditions. Systems that rely on foreign cloud infrastructure, lack telemetry isolation, or cannot be rolled back independently present unacceptable exposure in contested or classified environments. Operational governance considers whether the system can be paused, overridden, or re-tasked by sovereign authorities. It assesses the presence and effectiveness of human-in-the-loop or human-on-the-loop interfaces, the quality of behavioural logging during deployment, and the integration of mission specific fail-safes. AI systems deployed in battlefield or deterrence roles must be interruptible and traceable in real time, command control must not be undermined by inference opacity or procedural ambiguity.

Finally, legal and ethical governance metrics examine the capacity of sovereign institutions to understand, justify, and be accountable for AI decisions. These include traceability of outputs to legal actors, conformity with national and international humanitarian law, and the integration of audit frameworks across the AI lifecycle. A Sovereign AI system must be defensible not only in operation but in scrutiny by courts, legislatures, or coalition partners.

The Sovereignty Directorate would be responsible for defining thresholds and scoring scales within each of these dimensions. It would also maintain a sovereignty classification matrix, a tool to categorise AI systems by their assessed level of sovereign control across domains and functions. This matrix could be used to inform strategic investment decisions, determine assurance and validation priorities, and identify where sovereign development is essential versus where commercial or allied systems may be integrated under governance protocols.

These metrics are not static. They must evolve with technological change, adversary tactics, and operational complexity. Their development must involve continuous engagement with technical experts, legal authorities, operational commanders, and international partners. But without such metrics, the concept of sovereignty risks remaining rhetorical. With them, it becomes a testable, enforceable condition capable of informing strategy, enabling accountability, and preserving national authority in the age of automated decision-making.

7.2 Core Dimensions of Sovereign AI Metrics

The metrics should be multi-dimensional, reflecting both technical and institutional control. Each metric should be scored along a scale (e.g., low, partial, high assurance), calibrated to the mission context, risk profile, and operational domain. The proposed Directorate would use these to assess existing capabilities, inform procurement decisions, and guide resource allocation.

The assurance dimensions presented in this paper are ordinal and domain-specific by design. While scores such as “low,” “partial,” or “high” provide structured comparability across systems, they are not intended to be

aggregated into a single composite index. Sovereignty trade-offs are inherently contextual: a “high” requirement for legal auditability may outweigh “partial” assurance in compute control, depending on the mission domain, risk profile, and legal exposure. Future development of a formalised weighting or calculus, analogous to safety case frameworks in aviation or nuclear operations may support more rigorous trade-off modelling. However, in the current operational landscape, Sovereign AI assurance must remain a multi-factor, judgement-led process rooted in national values and strategic posture.

Data Sovereignty Metrics	• Percentage of training and inference data under national governance (classified, curated, stored) • Availability and rigour of data provenance audit • Assurance mechanisms for external data (validation, anomaly detection, red-teaming inputs)
Model Sovereignty Metrics	• Source control and access to model architecture, weights, and tuning logic • Auditability of model behaviour, including interpretability scores or mechanistic explanations • Rate of model reusability or modularity across secure domains
Training and Alignment Sovereignty	• Degree of in-house control over alignment objectives, reward functions, and tuning procedures • Proportion of models trained on sovereign compute with sovereign data • Traceability of training regimes, including documentation of strategic alignment goals

Compute Sovereignty

- Percentage of model training and inference conducted on nationally controlled infrastructure
- Hardware provenance scoring
- Security audit frequency and anomaly detection coverage in sovereign compute environments

Operational Governance Sovereignty

- Presence and effectiveness of override mechanisms (manual or automated)
- Degree of human-in-the-loop or human-on-the-loop integration in decision-critical functions
- Deployment traceability (audit trails, rollback capability, real-time behaviour logging)

Legal and Ethical Accountability

- Degree to which AI decisions can be traced to legally accountable actors
- Conformance with IHL, LOAC, and national legal standards

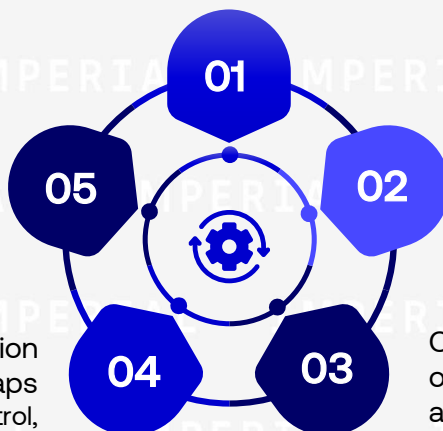
7.3 Implementation Role of the Directorate

The Defence AI Sovereignty Directorate would

Define scoring thresholds across these dimensions, tailored to mission criticality

Support alliance interoperability through mutual recognition frameworks for sovereignty scoring

Advise on R&D prioritisation based on observed gaps (e.g., weak compute control, opaque model logic)



Maintain a sovereignty classification matrix to triage AI systems by risk and assurance level

Conduct periodic reviews of systems in development and deployment

8. Refined Hypothesis for Mission-Driven, Modular Sovereignty

The central hypothesis of this white paper is that **the most strategically advantageous posture for any state seeking to maintain digital sovereignty is one of mission-driven, modular sovereignty. This approach asserts that sovereignty is not a binary status, but a multidimensional framework that must be applied selectively based on risk, function, and mission criticality.** It recognises that full sovereign control is essential in high-risk, mission-critical domains, including targeting, command support, cyber defence, battlefield autonomy, and information operations, where the absence of control introduces unacceptable strategic risks.

This hypothesis also acknowledges that selective collaboration in lower-risk domains can provide significant operational, economic, and strategic advantages. It allows states to leverage commercial innovations, benefit from trusted international partnerships, and access advanced technologies without compromising core security requirements. However, this collaboration must be carefully managed to prevent critical vulnerabilities,

including intellectual property theft, data exfiltration, and strategic dependency.

At its core, this hypothesis asserts that the path to effective AI sovereignty lies in the strategic differentiation of control levels, prioritising full sovereign command over high-risk systems, while maintaining flexibility in lower-risk areas. This approach supports the prioritisation of sovereign investments, the preservation of strategic autonomy, and the protection of national interests in a highly contested technological landscape.

However, achieving true sovereign control across these dimensions is a complex challenge, requiring careful navigation of several critical tensions. It must balance the operational necessity of control with the economic realities of cost, the strategic requirement for autonomy with the practical need for collaboration, and the ethical demands for accountability with the technical imperatives of innovation. This approach is consistent with the [AI Strategy \(s5.2.2\)](#), which emphasises the need for protecting critical technologies and selective intervention to ensure national security.

8.1 Rationale for the Hypothesis - UK MOD

The rationale for this hypothesis is grounded in the unique strategic, operational, and economic pressures facing the UK. It reflects

the need to evaluate and make decisions around trade-offs within several critical tensions:

Capability vs. Cost

Achieving full sovereign control across all AI systems is both resource-intensive and potentially cost-prohibitive. The MOD must prioritise its investments in Sovereign AI for those areas where the strategic risks of external dependency are highest, such as targeting systems, command decision support, and cyber defence. This aligns with the AI Strategy (s5.2.2), which emphasises the need to protect critical technologies, ensure onshore assured access, and safeguard UK intellectual property. It also recognises the importance of reducing hardware dependencies and ensuring secure, long-term access to critical AI infrastructure.

National Autonomy vs. International Influence

While strategic autonomy is essential for national security, it must be balanced against the need to integrate effectively with international partners. This requires trusted, secure data-sharing frameworks and modular interoperability that allow Sovereign AI systems to contribute to alliance operations without ceding control over core capabilities. This approach supports the UK's role within NATO, Five Eyes, and other critical alliances, while reinforcing national independence. It also reflects the AI Strategy's call for prioritising collaboration with like minded allies in areas where technology ubiquity poses minimal security risk.

Operational Flexibility vs. Industrial Prosperity

Domestic AI capability must not only meet operational requirements but also support broader economic resilience. This means aligning AI sovereignty with national industrial strategy, supporting local SMEs, and reducing reliance on foreign technology stacks. This approach reflects the AI Strategy's recognition of the need to selectively intervene to protect strategically important UK companies and capabilities from foreign influence. It also underscores the importance of maintaining a resilient domestic AI industry that can support long-term economic growth and technological leadership.

Strategic Independence vs. Technological Collaboration

While full sovereignty is critical in high-risk areas, maintaining technological leadership also requires collaboration with trusted partners. This includes co-development, joint research, and secure data exchange, aligning with the AI Strategy's emphasis on balanced technology protection and competitive advantage. This balance is essential for ensuring that the UK remains a leader in AI innovation while preserving the freedom to act independently when required.

8.2 Evaluation and Testing Pathways

For the hypothesis of mission driven, modular sovereignty to be operationally credible, it must be rigorously tested against real-world scenarios, diverse threat landscapes, and evolving technological risks. This process involves both structured experimentation and targeted stress-testing to validate the feasibility of Sovereign AI at scale.

The first phase of this evaluation should involve scenario based wargaming. This approach allows decision makers to test the

resilience of Sovereign AI architectures under realistic operational pressures. Wargaming can reveal critical vulnerabilities in AI systems, highlight integration challenges within joint and coalition frameworks, and provide empirical data to refine both technical designs and strategic doctrines. Scenarios should reflect the full spectrum of potential conflict, from grey-zone skirmishes to high intensity state-on-state warfare, and should incorporate both kinetic and non-kinetic elements.

These include

Grey-Zone Skirmishes

Conflicts in this category involve aggressive but ambiguous actions that fall below the threshold of conventional military response. Examples include the use of AI-enabled autonomous systems for surveillance and harassment in contested maritime zones, influence operations targeting digital public spaces, and coordinated cyber intrusions designed to degrade or disrupt critical

infrastructure. In such scenarios, AI systems must be capable of operating independently in environments where communications may be disrupted or contested, and where the political risk of escalation is high. They must also integrate seamlessly with human decision makers to ensure that tactical actions remain aligned with broader strategic goals.

Hybrid Warfare

This form of conflict blends conventional military force with irregular tactics, cyber operations, and information warfare. It is characterised by the simultaneous use of multiple domains to achieve strategic surprise or asymmetric advantage. For instance, a hybrid campaign might involve the use of AI-driven deepfakes to influence public opinion, coordinated cyberattacks

against military command networks, and the deployment of autonomous drones for kinetic strikes on critical infrastructure. Testing for hybrid warfare requires AI systems to demonstrate both resilience and adaptability, including the ability to rapidly switch between kinetic and non-kinetic modes of operation without external intervention.

High-Intensity State-on-State Warfare

These are large scale, conventional military conflicts across multiple domains, including land, sea, air, space, and cyberspace. In such scenarios, AI systems must support rapid decision making under conditions of extreme uncertainty, degraded communications, and contested control of the electromagnetic spectrum. They must also integrate with

legacy platforms and human command structures to enable joint and coalition operations at scale. This includes AI systems capable of coordinating massed fires, synchronising multi domain operations, and providing real-time situational awareness to dispersed units.

Kinetic Operations

These involve the direct use of force, including precision strikes, counter-air operations, and naval engagements. AI in this context must support target identification, strike coordination, and battle damage assessment, while maintaining compliance with the Law of Armed Conflict

(LOAC) and other legal frameworks. Systems must be robust against electronic warfare and capable of operating in denied environments where GPS, satellite links, and other critical data streams may be compromised.

Non-Kinetic Operations

These include cyber warfare, electronic warfare, and psychological operations aimed at degrading an adversary's ability to function without resorting to physical force. AI systems in this domain must be capable of detecting, disrupting, and defending against digital intrusions, while also supporting offensive operations such as

deep-packet inspection, spoofing, jamming, and network infiltration. They must also integrate with cognitive warfare capabilities, including the use of AI to influence enemy decision making through targeted information campaigns, algorithmic manipulation of digital platforms, and synthetic media generation.

Cognitive Operations

The use of AI to shape perceptions, influence behaviour, and manipulate the strategic calculus of adversaries. This includes the deployment of AI-driven propaganda, social media bots, and machine generated narratives designed to confuse, demoralise, or mislead opposing forces and their populations. It also involves the use of AI for sentiment analysis, predictive behavioural modelling, and real-time information environment monitoring to support strategic communications and influence operations.

Economic resilience must also be assessed through sensitivity analysis, evaluating the financial and industrial viability of Sovereign AI at different scales. This includes cost modelling for full spectrum vertical stacks versus more minimal, mission specific enclaves. It also requires an understanding of the broader economic impacts of Sovereign AI, including the potential to stimulate domestic industry, support high skill job creation, and reduce long term dependency on foreign technology.

Legal stress-testing is equally critical. This should involve a comprehensive assessment of the auditability and accountability frameworks required to ensure that AI

systems remain compliant with both UK and international law. Legal advisers, military commanders, and external ethicists should be engaged to map the risks associated with autonomous decision-making, AI-driven targeting, and command support systems. This process must also account for the unique legal challenges of operating in multinational coalitions, where differing legal regimes and accountability structures can complicate interoperability.

Finally, effective Sovereign AI requires robust alliance interoperability modelling.

This involves testing how Sovereign AI modules can be integrated into joint operations without compromising coalition efficiency or trust. It also means ensuring that UK developed AI systems can operate effectively within NATO, [Five Eyes](#), and [AUKUS](#) frameworks, while preserving full sovereign control over critical capabilities.

These evaluation and testing pathways are essential for validating the core hypothesis of mission-driven, modular sovereignty. They ensure that the UK can maintain operational independence while contributing effectively to collective security frameworks, reinforcing both national resilience and alliance credibility.

8.3 Strategic Deterrence, Legal Flexibility, and Adversarial Asymmetry

The legal and ethical governance of Sovereign AI must also be situated within the broader realities of strategic deterrence. The UK, like other democratic states, maintains capabilities such as nuclear and CBRN weapons that are governed under strict legal and political frameworks. These capabilities exist as credible deterrents under conditions of existential threat.

A parallel conversation is now emerging around artificial intelligence. As AI systems grow more capable and embedded in critical defence infrastructure, the question is no

longer simply how they function in tactical contexts, but how they may eventually shape or participate in strategic decision-making, up to and including deterrence, escalation control, and crisis response. **Sovereign AI governance must anticipate the possibility that AI systems could evolve into instruments of strategic leverage.** In such cases, states must be prepared not only to govern their use, but to defend their legitimacy in scenarios where adversaries may not share the same values, legal norms, or ethical constraints.

Historical precedent, most notably the UK's participation in the Manhattan Project, suggests that credible deterrent capability, combined with robust internal legal control, confers strategic influence. Preparation for AI's possible future role in deterrence should not be delayed until those capabilities are fully realised. It must begin now through legal foresight, sovereign design authority, and the development of assurance frameworks that retain lawful command even under asymmetric or existential threat.

Unlike nuclear weapons, AI systems are not governed by a single treaty regime. But international humanitarian law (IHL), the Law of Armed Conflict (LOAC), and emerging NATO doctrine all provide a basis for lawful AI deployment. Sovereign AI frameworks should therefore be designed with dual-use flexibility: constrained and governed for current operational use, but resilient and accountable under conditions of state-on-state escalation, where survival and strategic independence are at stake.

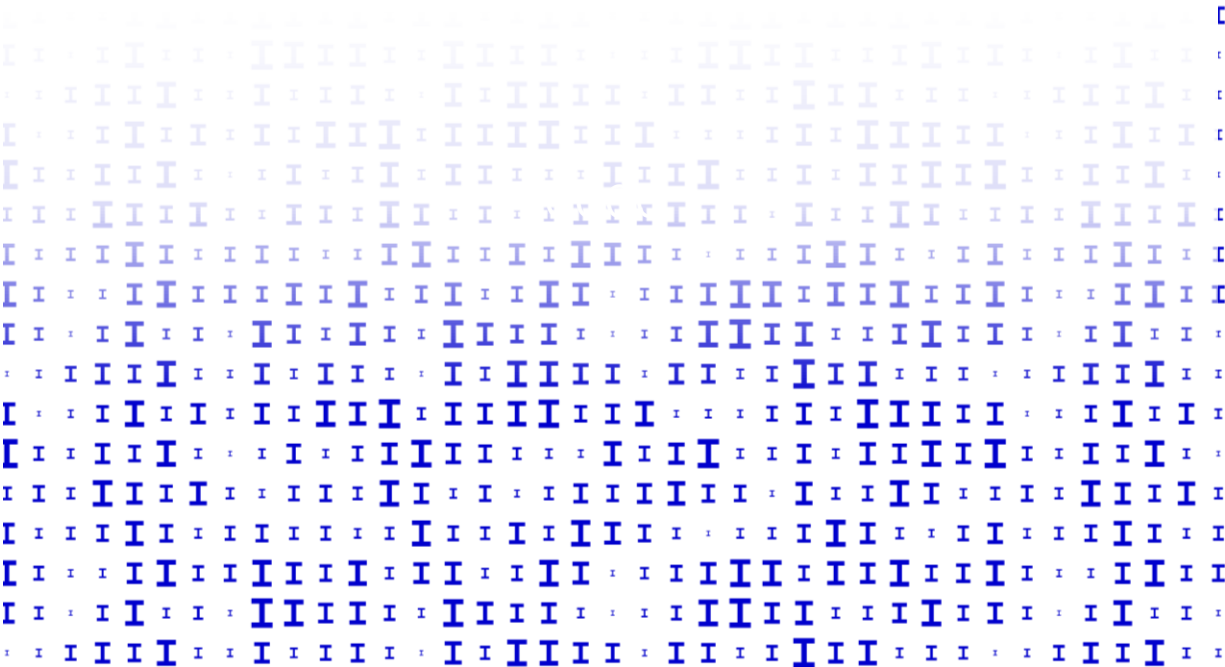
8.4 Forward Looking Utility

The mission-driven, modular sovereignty model presented in this paper offers more than a framework for current AI integration, it sets the foundation for long-term strategic advantage. It enables governments to concentrate sovereign resources where control is most critical to mission assurance, legal compliance, and strategic deterrence, while retaining the flexibility to adapt as technology, alliances, and adversaries evolve.

This approach ensures that Sovereign AI capability does not become a rigid doctrine, but a living posture, capable of withstanding pressure, scaling with operational need, and

aligning with broader legal and ethical commitments. It supports credible alliance contributions without compromising national command, and it preserves the authority to act independently when required.

Crucially, it positions states to lead in the emerging norms of responsible AI warfare, ensuring they are not just adopters of technology, but authors of its application and stewards of its legitimacy. In doing so, it future-proofs both operational readiness and national sovereignty in a world where digital power increasingly defines strategic freedom.



9. Risk Landscape and Adversarial Dependencies

The risks associated with dependence on non-Sovereign AI systems in defence are increasingly visible, and they span multiple layers of operational and strategic exposure. These risks are not only technical but legal, geopolitical, and doctrinal in nature.

Sovereign AI capability is not just about static assets or infrastructure, it is about agility. Rapid model adaptation, fine-tuning, red-teaming, and retraining must be exercised routinely as part of national preparedness. Major annual exercises should incorporate AI re-development cycles into planning and wargaming, engaging Defence Digital, DSTL, DAIC Connect partners, and cleared reservists with machine learning expertise. This “developmental velocity machine” should be treated as a capability in its own right; measured, tested, and refined across real-world constraints.

Non-sovereign models refer to AI systems developed, trained, or operated outside national control. Examples include commercial large language models accessed via public APIs, proprietary planning tools with undisclosed training data or objectives, or cloud-hosted AI services governed by foreign legal jurisdictions. These systems may lack transparency, update unpredictably, and be misaligned with domestic rules of engagement.

First and foremost, reliance on externally governed AI systems introduces the potential for loss of availability during moments of crisis. AI services hosted by commercial or foreign entities may be disrupted by policy changes, sanctions, or geopolitical realignments. In an alliance fracture scenario, a UK operated targeting or

cyber defence system that requires foreign authorisation or infrastructure access could become unusable or restricted.

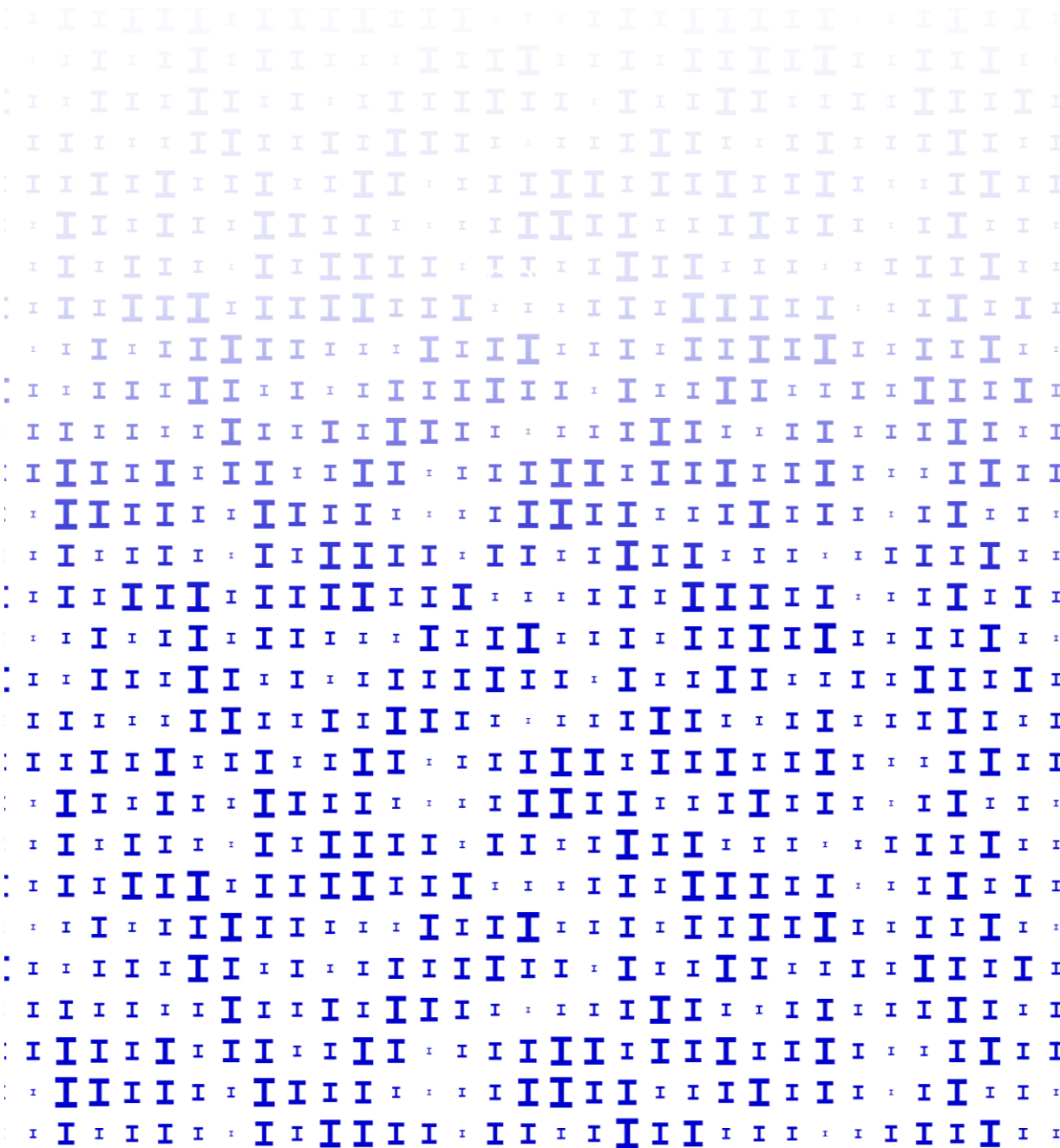
Second, there exists a legal discoverability risk. AI systems hosted or trained under foreign jurisdiction may be subject to litigation, disclosure demands, or surveillance requirements in those jurisdictions. This could expose sensitive MOD operational data, intelligence-derived training sets, or model inference logs to foreign scrutiny, undermining not only operational security but the UK’s sovereign decision-making integrity.

Third, and critically, non-Sovereign AI systems are vulnerable to model poisoning, prompt injection, data drift, and adversarial retraining attacks. These risks cannot be entirely eliminated, but they become significantly harder to detect or mitigate when model architecture and update control rest outside MOD’s own security perimeter. As AI-enabled threat actors increase their sophistication, the inability to retrain or patch mission-critical models on demand becomes a strategic liability.

[The Cyber Resilience Strategy For Defence \(2022\)](#) and the [National Cyber Strategy \(2022\)](#) both emphasise the imperative of building cyber capabilities that are resilient to disruption, adversarial interference, and system compromise. In high-consequence environments such as military command, targeting, and national infrastructure these strategies underline the importance of sovereign oversight, trusted supply chains, and assured digital architecture to ensure continuity of mission-critical functions under contested, degraded, or denied conditions.

Sovereignty is the key enabler of mitigation across these vectors. Sovereign AI systems can be air-gapped, independently audited, retrained in-theatre, and integrated with kinetic and digital feedback loops governed solely by MOD authorised staff. These capabilities cannot be purchased as off-the-shelf assurances, they must be built into the AI capability from design through to deployment.

Without Sovereign AI in high-risk operational domains, the MOD faces not only strategic risk, but a structural degradation of command assurance and operational continuity. The costs of compromise, whether through outage, espionage, legal challenge, or performance failure are not hypothetical. They are foreseen.



10. Comparative International Postures



The UK's choices on AI sovereignty are being made in a fast-moving international context. Other powers, both allied and adversarial have already begun reshaping their defence AI strategies around sovereignty, supply chain control, and political independence. Understanding these moves is essential to calibrating the UK's ambition, anticipating future interoperability challenges, and identifying strategic opportunity spaces.

The United States, while commercially dominant in AI development, is increasingly investing in digital sovereignty for its defence platforms. The Department of Defense has established the Chief Digital and Artificial Intelligence Office (CDAO) with a mandate to ensure interoperability, security, and explainability in all military AI applications. The Joint Warfighting Cloud Capability (JWCC) programme secures classified cloud and compute environments for training and deployment, ensuring that high-risk capabilities are not reliant on commercial infrastructure alone. While the US continues to export AI services to allies, it is reinforcing

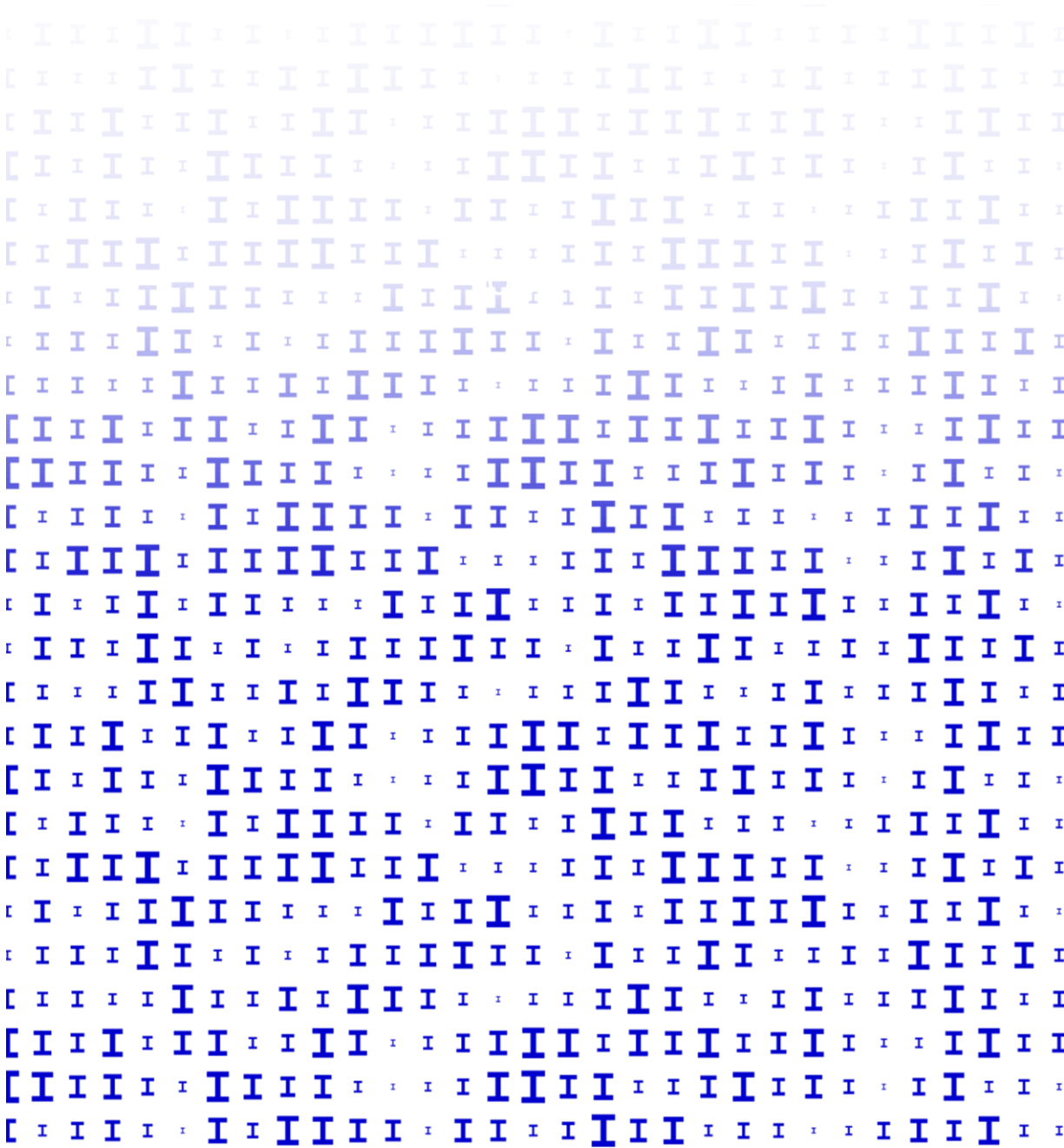
its core national systems with sovereign control.

France has taken a more explicitly sovereign path. Its Defence Artificial Intelligence Strategy commits to full lifecycle control of AI in mission-critical areas, including proprietary data labelling, on-premise model training, and embedded audit features in deployed AI agents. France's approach is deeply tied to its broader national strategy of technological autonomy, including national champions in both software and hardware domains.

Israel has adopted a vertically integrated model, coupling defence innovation hubs with frontline units and doctrine developers. Its AI systems are developed in tight coordination with operational commands, enabling rapid retraining, sovereign deployment, and direct integration with mission rules of engagement. Israel's model shows the effectiveness of a small, agile state embedding sovereignty not just in infrastructure but in organisational design.

China represents the most extreme case, having enshrined AI sovereignty in its national security and civil-military fusion doctrines. AI is treated not as a supporting tool but as a weapon of strategic dominance, with centralised governance over data, infrastructure, and model behaviour. While its governance model is incompatible with UK values, its ambition highlights the need for democratic states to maintain sovereignty as a precondition for responsible AI use.

The UK is uniquely positioned on the path that balances sovereignty with alliance integration, and legal accountability with technological agility. it can lead on AI interoperability standards, assurance frameworks, and ethical sovereign design by example. This will require investment, reform, and political commitment. But the alternative strategic drift and dependency would place the UK at a disadvantage not only on the battlefield, but at the negotiation table.



11. International Collaboration and Coalition Sovereignty

The pursuit of Sovereign AI capability by the UK must not be misunderstood as a retreat from alliance cooperation or a turn toward strategic autarky. On the contrary, sovereignty is the precondition for credible multilateralism. In an era of contested norms, accelerating technological diffusion, and multipolar security architectures, sovereign control over AI systems enhances the UK's capacity to operate with trusted partners, shape emerging standards, and contribute responsibly to joint force structures. It ensures that interoperability is not conditional on dependence, and that UK forces remain both coalition-ready and command-secure.

As outlined in Chapter 3, the concept of **non-compromising interoperability** must serve as the foundation for AI-enabled alliance operations. Sovereignty and interoperability are not in opposition; they are dual requirements for legitimacy and effectiveness in digitally enabled coalitions. The future of multinational operations will depend on the ability of allied systems to

exchange data, coordinate action, and align objectives without forfeiting legal authority, operational control, or doctrinal consistency. This requires the deliberate development of what this paper terms coalition sovereignty: a model in which nations retain exclusive control over the core functions, parameters, and governance of their AI systems, while enabling modular, standards-based cooperation across defined technical and operational interfaces.

The UK is well positioned to lead this effort. Its participation in key multilateral structures such as NATO, Five Eyes, AUKUS, and the Joint Expeditionary Force provides a robust framework for engagement. Its credibility as a responsible AI actor, grounded in rule-of-law traditions and high-ethics defence doctrine, enables it to shape normative standards with legitimacy. And its technical ecosystem, spanning Defence Digital, DSTL, the Defence AI Centre, and academic partners offers a strong foundation for both conceptual leadership and technical experimentation.

11.1 Multilateral Structures as Vehicles for AI Norm-Setting

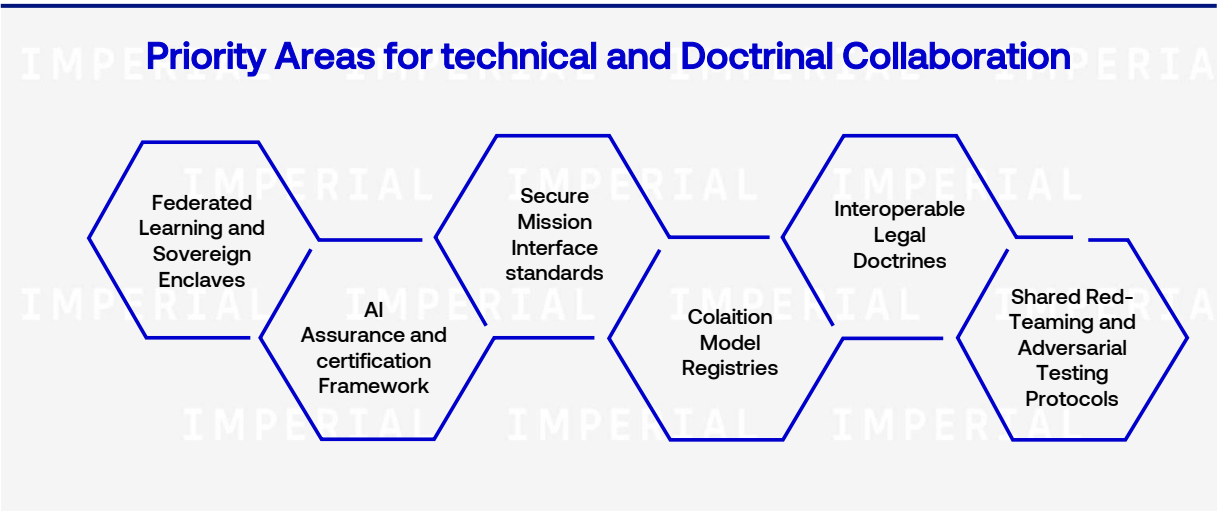
Several existing alliances and frameworks, introduced earlier in this paper, provide natural pathways for advancing shared standards on sovereign and interoperable AI. These multilateral platforms should not be viewed merely as diplomatic forums or policy coordination tools. Rather, they are essential operational laboratories, spaces where trust based, sovereign-aligned, and legally defensible models of AI-enabled coalition warfare can be prototyped,

validated, and scaled. The UK's active engagement in these mechanisms, already reflected in doctrine and policy as shown throughout this paper, must now be matched with a strategic mandate to institutionalise Sovereign AI governance at the coalition level. What follows is a structured articulation of how these platforms can now be used to formalise a model of strategic coordination we term **coalition sovereignty**.

- [NATO's Artificial Intelligence Strategy \(2021\)](#) outlines six guiding principles: lawfulness, responsibility, explainability, reliability, governability, and bias mitigation. The UK has a leadership opportunity to help translate these high-level principles into enforceable certification frameworks that define minimum sovereign thresholds for AI used in collective defence. As discussed earlier, NATO is also advancing concepts of trusted interfaces and federated AI infrastructure that complement the UK's strategic emphasis on layered sovereignty.
- [The Tallinn Manual](#) and NATO CCDCOE serve as platforms for shaping legal norms around AI and autonomy. These forums provide the procedural foundation for aligning interpretations of International Humanitarian Law (IHL) and legal accountability frameworks across Sovereign AI deployments. The UK should use these channels to codify legal interoperability mechanisms, particularly for AI-enabled decision support and autonomous engagement systems.
- The Five Eyes intelligence alliance, referenced earlier, has long provided a platform for deep coordination of technical capabilities, especially in signals intelligence and cyber defence. Its trusted nature makes it ideally suited for sensitive work on joint model validation tools, AI threat sharing protocols, and secure red-teaming environments, particularly in domains like ISR fusion and cyber operations.
- AUKUS, also introduced earlier as a testbed for autonomous systems interoperability, includes AI and autonomy as flagship areas of Pillar II. It represents a forward leaning opportunity for co-development of sovereign-by-design systems, shared deployment architectures, and mission specific federated inference tools. The UK, US, and Australia are already engaged in operational experimentation under AUKUS auspices, making this an ideal arena for advancing shared assurance metrics and ethical operational frameworks.

11.2 Priority Areas for Technical and Doctrinal Collaboration

To ensure that Sovereign AI systems can integrate effectively into allied operations while preserving UK legal and operational control, this paper identifies six core areas for targeted international collaboration:



AI Assurance and Certification Frameworks

01

The development of common audit and certification protocols across allies is essential to build trust in Sovereign AI systems. These frameworks should include mechanisms for model validation, dataset lineage tracing, inference logging, and update transparency enabling each nation to independently verify that coalition partners' systems meet mutually agreed standards without requiring access to classified internals.

Federated Learning and Sovereign Enclaves

02

Allies should jointly develop federated learning architectures that allow national models to be trained on shared operational datasets such as ISR feeds or cyber telemetry without exposing raw data or model internals. These architectures should support sovereign enclaves that allow each partner to maintain model control while benefiting from coalition-wide data diversity and scenario generalisation.

Shared Red-Teaming and Adversarial Testing Protocols

03

AI-enabled red-teaming, already essential for sovereign assurance, should be extended to coalition exercises. Allies should develop joint simulation environments, threat libraries, and stress-testing frameworks that expose shared vulnerabilities in autonomous systems, model inference chains, and human-machine decision loops.

Interoperable Legal Doctrines

04

The UK should lead in aligning interpretations of International Humanitarian Law (IHL) as applied to AI-enabled systems, especially in domains like autonomous targeting, battlefield inference, and command decision support. This includes establishing shared definitions of meaningful human control, as well as reciprocal accountability structures that clarify responsibility in joint operations involving AI.

Coalition Model Registries

05

Where appropriate, the UK should advocate for the creation of classified coalition model registries and shared repositories that track the operational deployment, validation status, and doctrinal integration of AI models used in allied missions. Access controls and national flags would preserve sovereignty while enabling trust-based integration of verifiably governable systems.

Secure Mission Interface Standards

06

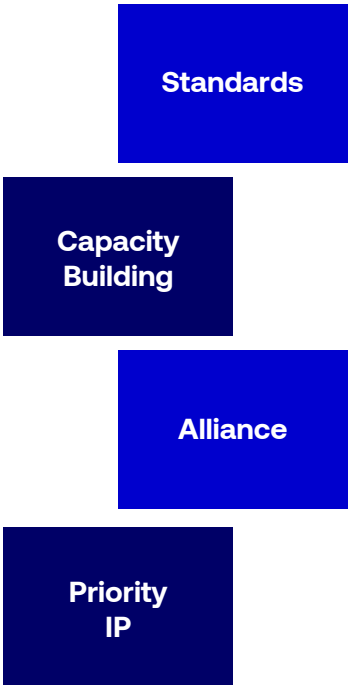
Building on existing NATO interoperability protocols (e.g. [STANAGs](#)), the UK should help define interface standards for AI systems in coalition operations. These should ensure that data, inferences, and recommended actions can be exchanged securely and reliably without exposing core model behaviour or compromising domestic oversight structures.

11.3 Diplomatic and Strategic Considerations

Pursuing international collaboration on Sovereign AI must be underpinned by a clear understanding of the strategic risks and dependencies it seeks to mitigate. As illustrated, reliance on private infrastructure, such as Starlink, without sovereign guarantees has exposed operational vulnerabilities and raised questions of accountability. Similar risks arise if coalition AI systems are dependent on vendors or platforms whose legal, ethical, or political commitments diverge under pressure.

The UK must therefore ensure that international collaboration is rooted in sovereign assurance, not commercial convenience.

Partnership must not substitute for control; rather, it must be built on verifiable autonomy and reciprocal transparency. Engagement should be in technological diplomacy with like-minded nations, beyond core alliances, to shape a shared normative environment for military AI. Engagements with EU partners, Indo-Pacific democracies, and members of the OECD AI Network can help build a broader coalition of trusted states committed to the responsible development and use of AI in security domains. Sovereign AI does not preclude collaboration but it does require clarity on the terms, thresholds, and structures through which collaboration occurs. States must navigate a spectrum of partnership models, from informal knowledge sharing to tightly governed technology integration. Strategic AI collaboration should be guided by the following pillars:



Standards

Establishing shared technical, ethical, and legal standards for Sovereign AI is foundational. These standards should cover data provenance, model auditability, system override, and accountability thresholds. Common frameworks whether through NATO, AUKUS, or multilateral accords can facilitate interoperability without compromising sovereign integrity.

Capacity Building

Defence partnerships should invest in developing Sovereign AI capabilities across allied nations, particularly through joint training programmes, red-teaming exchanges, and shared assurance tooling. Building collective resilience strengthens alliance readiness while preserving national control.

Alliances

AI cooperation must reflect the strategic depth and political trust of the alliance it operates within. High-trust relationships (e.g., Five Eyes, AUKUS) may permit deeper integration and shared assurance regimes, while looser coalitions may limit cooperation to interface level compatibility or federated learning models. The nature of the alliance defines the level of acceptable risk and shared control.

Priority IP with Conditional Export

Certain sovereign capabilities, especially foundational models, secure inference engines, or validated safety tooling may be shared or exported under strict conditions. This should be managed through controlled licensing, partner vetting, and export controls to ensure that strategic IP is not diluted or misaligned in downstream use. Controlled exports can strengthen collective posture while retaining national advantage.

11.4 From Coalitions of Convenience to Coalitions of Sovereignty

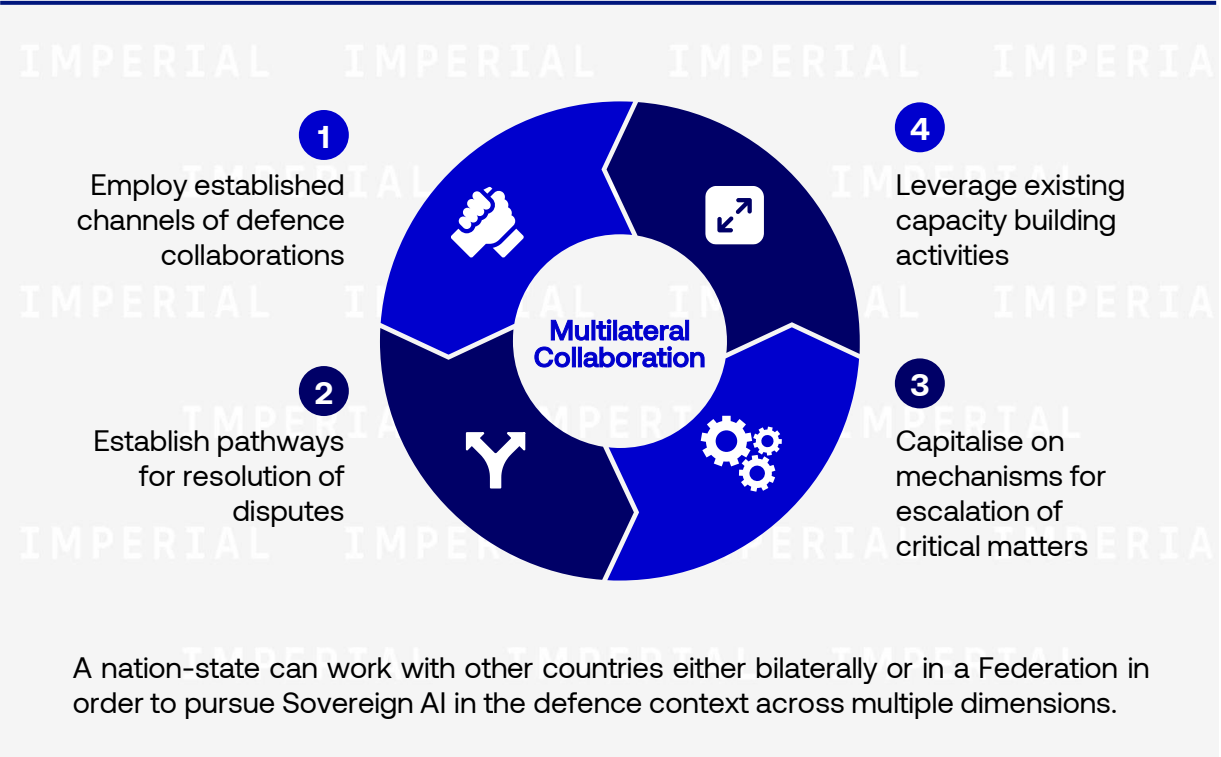
The future of AI-enabled warfare will not be defined solely by national capability, but by the ability to integrate those capabilities into credible, lawful, and strategically coherent coalitions. Nations that can build trust, through assurance, transparency, and control, will shape the standards by which AI is governed in conflict. Nations that cannot find themselves constrained by systems they do not fully understand or control.

By embedding Sovereign AI into its alliance posture, the UK can ensure that future coalitions are not merely interoperable but co-sovereign, capable of acting together without compromising command, legality, or national judgement. This model of coalition sovereignty is not a compromise between independence and integration; it is the only viable framework for trusted collective security in the digital age.

11.5 Leveraging Partnerships and Multilateral Collaboration

The UK’s ability to shape Sovereign AI norms and alliance architectures depends not only on infrastructure and policy, but on strategic contributions that confer influence. The UK is home to key talent pools, hosts critical AI research centres, and plays a convening role across Five Eyes, NATO, and AUKUS. Many leading AI labs have UK-based researchers and partnerships. However, this influence will not materialise passively. It requires

investment in sovereign capability, secure domestic infrastructure, and institutional leadership so that UK developed components, datasets, and alignment pipelines become foundational to interoperable allied systems. Without credible contributions, the UK risks becoming a consumer of strategy rather than a co-author of future norms.



1) Employ Established Channels of Defence Collaborations

Nation-states seeking to advance Sovereign AI capabilities in defence can strategically utilise pre-existing defence collaboration frameworks to accelerate joint development, ensure interoperability, and foster trust. Bilateral alliances such as those underpinned by mutual defence agreements provide a strong foundation for co-developing Sovereign AI systems that reflect shared security priorities and ethical standards. Similarly, multilateral frameworks like NATO,

AUKUS, the Five Eyes and the European Defence Agency offer structured environments where partners can exchange data, co-invest in AI-enabled military technologies, and develop shared protocols for the use of autonomous systems. These established channels reduce the barriers to trust and integration, allowing Sovereign AI to be embedded within collective defence postures while maintaining national strategic autonomy.

2) Leverage Existing Capacity-Building Activities

Collaborative approaches to Sovereign AI should build upon ongoing capacity-building efforts in defence technology and digital infrastructure. These may include joint research programmes, officer exchange schemes, war-gaming exercises, and multinational AI testbeds. Leveraging these initiatives enables countries to strengthen their technical expertise, institutional readiness, and doctrinal familiarity with AI tools. By embedding Sovereign AI

development within existing education and training partnerships, nations can align capabilities incrementally, reducing duplication while tailoring solutions to distinct national contexts. This approach also allows smaller or less technologically advanced states within a federation or alliance to participate meaningfully in Sovereign AI ecosystems through tiered development and shared access to common resources.

3) Capitalise on Mechanisms for Escalation of Critical Matters

In the rapidly evolving and high-stakes domain of AI in defence, partner nations must maintain clear and robust mechanisms for the escalation of critical matters. This includes agreed protocols for raising concerns related to AI model behaviour, data security breaches, or the failure of AI-enabled systems during joint operations. Structured escalation pathways, ranging from diplomatic consultations to joint military AI oversight boards, can help mitigate

misunderstandings, prevent unilateral actions, and preserve the integrity of shared missions. Escalation mechanisms should be codified in formal agreements or memoranda of understanding and aligned with each nation's legal and operational doctrines. These tools are especially vital in federated AI systems, where decentralised nodes operate with varying levels of control and visibility across different jurisdictions.

4) Establish Pathways for Resolution of Disputes

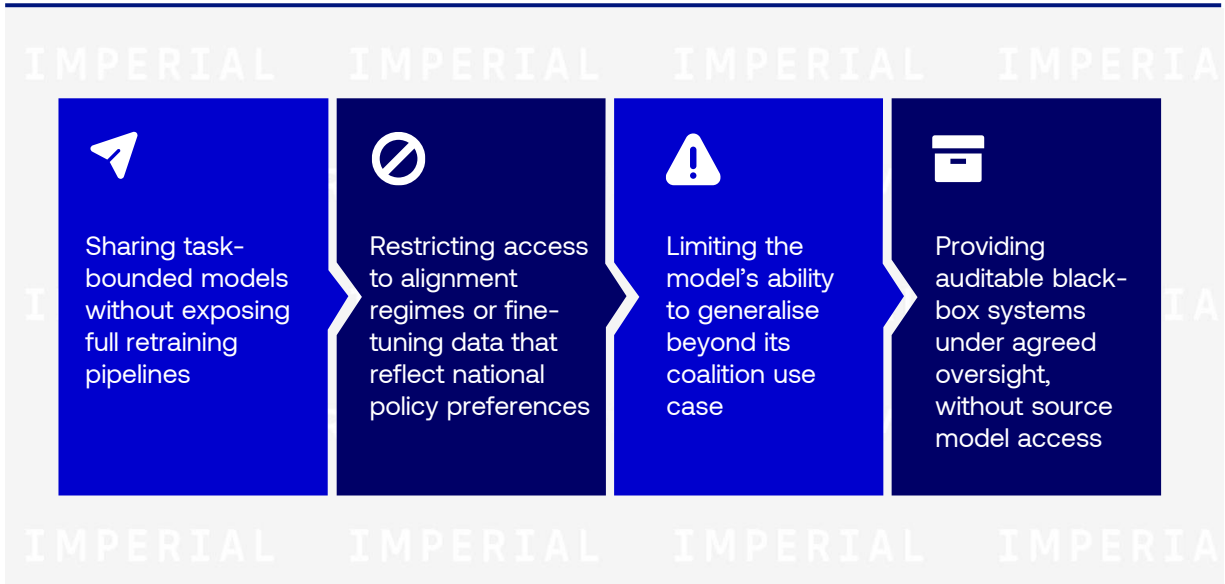
Effective dispute resolution is essential for sustaining long-term AI cooperation in defence, particularly where national sovereignty and security imperatives intersect. Countries must agree on transparent, equitable mechanisms for resolving disagreements over issues such as data ownership, attribution of AI errors, intellectual property in joint developments, and the operational deployment of shared systems. This could involve standing arbitration panels, recourse to international

legal forums, or specially mandated AI ethics committees. Importantly, these pathways should balance the protection of sovereign interests with the need for predictable and rules-based collaboration. Institutionalising such mechanisms within defence treaties or federated AI governance frameworks reduces the risk of breakdowns in trust and ensures that Sovereign AI development remains accountable, inclusive, and strategically aligned.

11.6 Export Controls and Technological Assurance in Sovereign AI

As Sovereign AI capabilities become embedded in critical defence functions, the question of how and when to share these technologies with international partners demands careful consideration. While AI collaboration is essential for coalition operations, standardisation, and trusted interoperability, not all capabilities can or should be exported in their entirety. Some models, datasets, and system behaviours, particularly those involving mission-specific logic, override architecture, or security response patterns may introduce strategic

risk if transferred without constraint. To manage this, states may need to adopt a model of differentiated technical release, whereby AI capabilities are selectively scoped, tailored, or compartmentalised prior to integration with partner systems. This practice is well established in other sensitive domains, such as radar systems, encryption modules, or electronic warfare suites where export versions are adjusted to meet alliance requirements without compromising sovereign integrity. In the AI context, this might involve:



As Sovereign AI capabilities mature, the UK and its allies will increasingly face decisions on how and when to share AI systems with international partners. To manage this, the UK should develop export control policies specific to AI models and systems, particularly for applications with strategic or dual-use potential. This may include the creation of tiered or “controlled-release” versions of AI capabilities, akin to how military platforms are sometimes exported in downgraded configurations. Such an approach would allow the UK to support trusted partners with AI-enabled tools while safeguarding sensitive capabilities, model architectures, and training datasets that could pose national security risks if widely

proliferated. Embedding export governance into AI procurement and assurance frameworks will ensure that sovereignty extends not only to the use of AI but also to its dissemination. Such practices will require new export classification schemes, legal frameworks, and technical architectures to support collaboration without dilution of sovereign control. Importantly, they also enable the UK and its partners to participate in alliance operations from a position of trust where capabilities can be verified, integrated, and jointly governed, but not repurposed in ways that could jeopardise national security.

12. Conclusions

Artificial Intelligence now sits at the core of modern military capability. It enables faster decision making, enhances threat detection, and increases the autonomy and precision of deployed forces. But with these benefits come profound questions of accountability, legality, and control. The United Kingdom, as a democratic military power with global reach and high legal standards, must ensure that it retains the authority and technical capability to govern the AI systems it relies upon.

This white paper has presented the case for a posture of targeted AI sovereignty: a calibrated approach where sovereign control is asserted in domains of high operational and legal risk, and selectively relaxed in areas where commercial or allied integration poses minimal strategic exposure. It has shown that such a posture is achievable, fiscally, industrially, and strategically, if pursued with intent, discipline, and clear ownership.

12.1 Implications for Policy and Capability Development

A Sovereign AI posture demands action at multiple levels. Procurement processes must include sovereignty thresholds. Legal and ethical assurance functions must be embedded into AI capability lifecycles. Infrastructure must be secured to enable sovereign model training and deployment. Personnel policies must build a workforce fluent in both the technical and doctrinal dimensions of AI governance.

At the governance level, the MOD must move from concept to execution. A permanent directorate or unit, mandated to coordinate AI assurance, policy integration, and technical capability management must be established. Existing structures such as the Defence AI Centre and Defence Digital must be aligned to deliver on sovereign capability priorities, with cross-agency links to the Office for AI, the NCF, and DSIS delivery programmes.

12.2 Strategic Imperatives and the Cost of Inaction

Artificial intelligence now structures how states perceive threats, make decisions, deploy force, and respond under pressure. As such, control over defence AI systems is not a matter of technological preference. It is a condition of sovereignty. This paper has argued that such control, achieved through targeted, layered, and mission-driven sovereignty, is no longer optional. It is a requirement for credible military authority, lawful force projection, and strategic freedom of action in an era of accelerating automation.

The systems being adopted today, models,

datasets, decision pipelines, platforms, and processors, are embedding assumptions that will shape command behaviour. They will determine escalation thresholds, targeting norms, coalition dependencies, and legal exposure. The architecture of future operations is being laid now, and once scaled, these systems will not be easily unwound or reformed. **States that do not act today will find themselves not only technologically dependent, but strategically encumbered, governed by tools they cannot fully explain, modify, or ethically justify in moments of consequence.**

The cost of inaction is not speculative. It is the quiet erosion of command authority, the weakening of alliance trust, and the loss of institutional confidence when decisions are driven by systems developed elsewhere, aligned with priorities not their own. It is the legal risk of deploying systems that cannot be audited. It is the operational risk of failure under contested conditions. And it is the political risk of being unable to account for decisions of force that have escaped institutional review.

To avoid this, states must now act decisively. They must designate Sovereign AI as a strategic capability class, one subject to the same levels of oversight, assurance, and accountability as nuclear command, intelligence collection, or kinetic targeting. This requires standing institutions, not transient programmes. It demands that procurement policy reflect sovereignty thresholds, that red-teaming and legal audit be embedded by design, and that alliance engagement be conducted from a position of assured control, not inherited reliance.

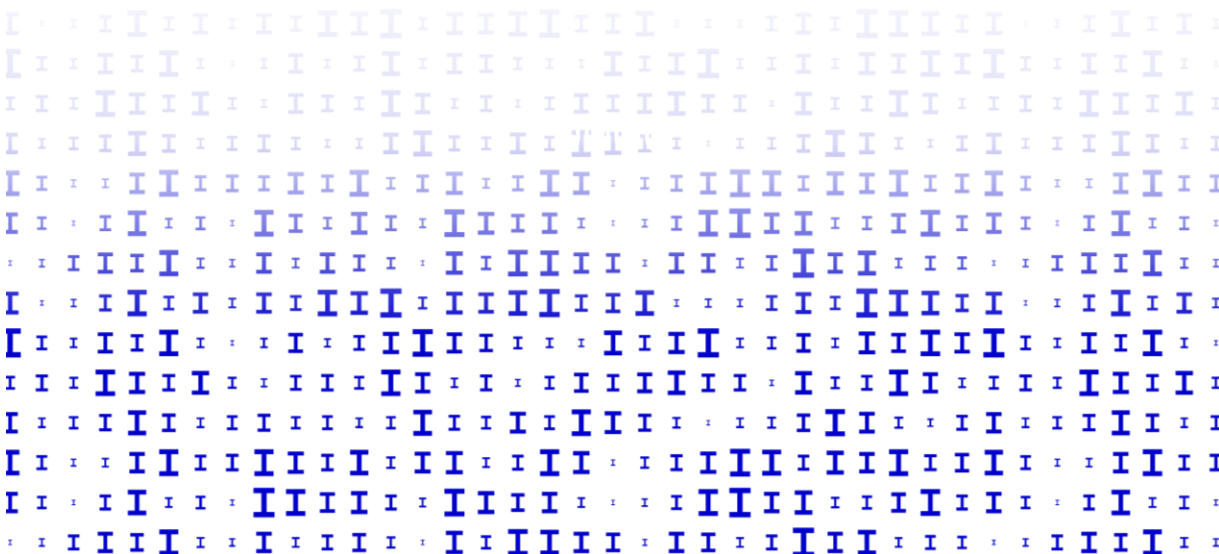
States that lead will shape the standards by which AI is developed, deployed, and governed in military contexts. They will define what it means to use automated systems lawfully, ethically, and safely. They

will retain the ability to act under their own rules, using their own systems, with the confidence that decisions taken in extremis will be traceable, defensible, and sovereign.

States that fail to lead will operate under architectures built by others. They will face strategic crises with capabilities they do not fully control. They will interpret the world through models they did not build, using platforms they cannot secure, shaped by values they did not define. This is not a moment for incrementalism. It is a moment for structural commitment. Sovereign AI must become an organising principle in defence planning, capability development, legal doctrine, and international posture. Anything less, risks the forfeiture of strategic agency in a domain that is fast becoming the nervous system of military power. Posture must now become policy. States that define and institutionalise their AI sovereignty posture today will shape the rules of digital-era conflict tomorrow.

The judgement required is not whether to act, but how quickly and with what clarity of intent.

The age of automated conflict has begun. The authority to lead within it will belong only to those who have retained the authority to decide.



About the Authors

[Adele Jashari](#) is the Deputy Director of Translational Research for the Trusted AI Alliance, advises several AI companies and was part of the team that created the [Commonwealth Fintech Toolkit](#). She brings extensive experience in leveraging emerging technologies to drive innovation and growth across industries. Adele began her career in finance working in the city for leading investment banks such as Goldman Sachs and Société Générale.

[David Shrier](#) is a Professor of Practice, AI and Innovation, with Imperial College London, where he leads the [Trusted AI Alliance](#) (a multi-university collaborative focused on building responsible & trustworthy AI). Initiatives he pioneered for Imperial, MIT and University of Oxford delivered more than US\$ 1 billion of financial support. He has helped over 100 governments to develop technology policy including advising the European Parliament on the [EU AI Act](#), and leading the team that created the [Commonwealth Fintech Toolkit](#). Other relevant work includes creating Oxford's [Leadership and Diversity for Regulators](#) programme, helping more than 50 nations shape financial inclusion policy. His latest book, [Basic AI: A Human Guide to Artificial Intelligence](#) was published in 2024. In his private sector activities, he has led more than \$1bn of technology-enabled growth initiatives; as CEO conducted an oversubscribed initial public offering on the New York Stock Exchange; and through his venture studio [Visionary Future](#) and other private equity, has invested in more than 50

enterprises, primarily in the AI domain.

[Aldo Faisal](#) holds a prestigious UKRI Turing AI Fellowship, is a Professor of AI at Imperial College London and holds the Chair in Digital Health at the Universität Bayreuth (Germany). He is since 2019 the founding director of the £50 million UKRI centres in [AI for Healthcare](#) and [AI for Digital Health](#) in London. Since 2024 he is Director of Science and Innovation at the Alan Turing Institute responsible for the Grand Challenge in Health, the national AI research institute of the United Kingdom. His research focusing on Generative AI for health and Human-AI interfacing has won numerous international prizes and in 2024 the Federal German Government and Parliament appointed him as member of the German Ethics Council. He is an author of the Amazon Global Top 10 textbook *Mathematics for Machine Learning*, and technical capability to govern the AI systems it relies upon.

This white paper has presented the case for a posture of targeted AI sovereignty: a calibrated approach where sovereign control is asserted in domains of high operational and legal risk, and selectively relaxed in areas where commercial or allied integration poses minimal strategic exposure. It has shown that such a posture is achievable, fiscally, industrially, and strategically, if pursued with intent, discipline, and clear ownership.

This paper was written by human authors, augmented by LLMs.



Glossary of Terms and Abbreviations

AI Assurance - The structured validation and verification of AI systems to ensure that they behave as intended, under operational conditions, and within acceptable risk and safety bounds. This includes robustness testing, interpretability, red-teaming, and legal auditability.

AI Life Cycle - The end-to-end process of AI system development and management, including data acquisition, model training, validation, deployment, continuous learning, updating, and decommissioning.

Artificial Intelligence - The field of computer science concerned with building systems that can perform tasks typically requiring human intelligence such as perception, reasoning, learning, and decision-making. In defence, AI is increasingly embedded in ISR, targeting, logistics, and C2 workflows.

Alignment (AI Alignment) - The process of ensuring that the objectives, behaviours, and outputs of an AI system remain consistent with human intent, strategic policy, and legal constraints.

Autonomy - The capacity of a system to perform tasks or make decisions without direct human intervention. Tactical autonomy refers specifically to battlefield-relevant functions such as navigation, target identification, and engagement conducted by AI-enabled platforms.

C2 (Command and Control) - The exercise of authority and direction by a properly designated commander over assigned and attached forces in the accomplishment of a mission.

COA (Course of Action) - A military planning term referring to a potential operational plan developed and evaluated to achieve specific objectives under defined constraints.

Cognitive Security - The protection of decision-making environments, populations, and institutions from manipulation through disinformation, influence operations, or adversarial AI-generated content.

Compute Sovereignty - The ability to govern where and how AI models are trained and executed, including control over hardware, cloud infrastructure, and physical data centres. Essential to safeguarding inference integrity and data confidentiality.

Cyber Operations - Encompasses both defensive and offensive actions taken in cyberspace to protect, disrupt, degrade, or influence digital systems and infrastructure.

Data Sovereignty - The legal and operational control over data used to train, validate, and operate AI systems, including data origin, curation, classification, and access governance.

Defensive Cyber - The use of AI and automation to monitor, detect, and respond to malicious cyber activity targeting national defence infrastructure.

Ensemble Assurance - The process of validating the combined behaviour of multiple AI systems operating in coordination, particularly relevant in autonomous vehicles, sensor networks, and multi-domain integration.

Fail-Safe Architecture - The design of AI-enabled systems to safely revert, suspend, or hand back control to human operators when anomalies, legal constraints, or operational risks are detected.

Governance Sovereignty - The capacity to define and enforce policies, protocols, and legal frameworks governing AI use in military operations, including the ability to audit, override, or suspend AI behaviour as needed.

HIL / HOTL (Human-in-the-Loop / Human-on-the-Loop) - Concepts referring to the integration of human oversight in automated systems. HIL ensures humans approve every action; HOTL allows systems to act autonomously under human supervision with override capability.

IHL – International Humanitarian Law - A body of legal rules which regulates the conduct of armed conflict and protects persons not participating in hostilities. AI-enabled systems used in warfare must comply with IHL principles of distinction, proportionality, and precaution.

Institutional Capability - The personnel, expertise, and organisational structures required to govern, validate, and adapt Sovereign AI systems across mission environments.

Intelligence, Surveillance, and Reconnaissance (ISR) - A critical defence function in which AI is increasingly used to collect, process, and fuse data from multiple sources to generate operational intelligence and inform targeting or threat assessments.

Legal and Ethical Sovereignty - The ability to ensure that AI systems operate within national legal frameworks and ethical norms, with mechanisms for legal review, traceability, and public accountability.

Lifecycle Governance - Oversight across the entire AI system lifecycle—from data curation and training, to deployment, retraining, and decommissioning—ensuring traceability and sovereign control.

LOAC (Law of Armed Conflict) - Closely related to IHL, LOAC encompasses international agreements and customary law governing military engagement, including state responsibility for automated decisions.

Mission-Driven Sovereignty - A posture in which sovereignty is prioritised based on the strategic importance, legal exposure, and operational consequences of a given AI application. Rejects autarky in favor of calibrated, risk-informed control.

Model Sovereignty - The capacity to design, inspect, modify, and validate AI models, including architecture, weights, objectives, and tuning parameters. Ensures systems remain aligned with national doctrine and strategic priorities.

Narrative Battlespace - The contested information environment in which states and non-state actors compete to shape perceptions, control narratives, and influence behaviour through digital, media, and psychological means.

Non-Sovereign Models - AI systems developed or hosted externally—such as commercial platforms (e.g., GPT-4, Google Vertex AI, Amazon SageMaker)—for which the state lacks access to source code, control over updates, or legal accountability structures.

Offensive Cyber - AI-enabled or supported cyber operations designed to disrupt, degrade, or influence adversary digital assets, networks, or capabilities, often covertly or under conditions of strategic ambiguity.

Operational Governance - Oversight mechanisms that ensure deployed AI systems behave within defined legal, strategic, and operational parameters, including real-time audit logging, failsafe interventions, and decision traceability.

Red-Teaming - A structured method of testing AI systems by simulating adversarial attacks, failures, or edge cases to identify vulnerabilities and ensure robust, lawful operation.

Sovereign AI - AI systems that are governed, operated, and assured under national control, across all six dimensions of sovereignty: data, model, training and alignment, compute, operational governance, and legal and ethical accountability.

System-of-Systems - A configuration of independent systems, both human and machine, that function collectively to achieve complex mission objectives. Requires coordination, interoperability, and sovereign control of each component's behaviour.

Targeted Sovereignty - An approach that focuses sovereign investment on AI systems used in high-risk, high-consequence mission

areas, while allowing for commercial or collaborative solutions in lower-risk domains under strict governance.

Telemetry - Raw operational data transmitted from deployed systems, including sensor logs, network activity, and system diagnostics often used as inputs for model training or anomaly detect.

