ATTACKS TO RESILIENCE: AI IN CRITICAL INFRASTRUCTURE

Exploring security risks from adversarial AI and AI adoption in UK critical national infrastructure.

AUGUST 2026

Microsoft

plexal

ABOUT THIS PAPER

Artificial intelligence is increasingly embedded in the systems that underpin critical national infrastructure (CNI), delivering benefits across UK public services, from medical diagnostics to fraud detection[1]. While these early successes highlight AI's potential to improve efficiency, they also introduce two key challenges for CNI providers:

1. Security risks associated with AI adoption, particularly in the integration of information technology (IT) and operational technology (OT) systems.

2. Threats from adversarial actors targeting AI systems within CNI.

Recent research has outlined how AI systems can be attacked; less attention has been paid to how these risks manifest in operational environments, and how organisations can build resilience in response.

The UK's National Security Centre (NCSC) framework on adversarial machine learning provides a technical taxonomy of these attack types[2]. This paper, developed by Plexal and Microsoft with support from the Laboratory for AI Security Research (LASR), builds on that foundation to examine the challenges within CNI, focusing in particular, on the energy, water, transport, telecoms, healthcare, finance, and emergency services sectors (see appendix for full CNI categorisation), and sets out strategic recommendations relevant across operators and solution providers.

ABOUT LASR

The Laboratory for AI Security Research (LASR) is a collaboration between the public and private sectors in the UK to bring together the best minds in AI security. LASR is

dedicated to mitigating security risks to and from artificial intelligence (AI) to strengthen national security and support economic growth.

Launched in November 2024 at the NATO Cyber Defence Conference, the initiative brings together world-leading experts from UK organisations including Plexal, University of Oxford, The Alan Turing Institute, Queen's University Belfast and the UK Government, alongside a broad network of academic, industry, and international partners.

LASR conducts cutting-edge research at the intersection of AI and cyber security, develops novel capabilities and skills, accelerates research commercialisation, and fosters international collaboration for the secure development and deployment of AI.

This work was authored by Plexal, and supported by Microsoft and LASR. The views expressed in this paper are those of the authors and do not necessarily reflect the position of Microsoft, LASR or His Majesty's Government.

AI IN CNI: OPPORTUNITIES AND RISKS

Leveraging AI in diverse domains, from cloud networks to industrial systems, can increase efficiencies and improve cyber defences. At the same time, it can also amplify cyber security risks and reshape how those risks emerge, spread, and scale. As AI connects IT and OT (operational technology, like industrial systems), it creates more efficient but more fragile environments where failures, attacks, or errors can propagate rapidly across systems.

CNI is already under sustained pressure. The US Cybersecurity & Infrastructure Security Agency[3] has shown how state-sponsored groups are exploiting unpatched telecoms and network devices to gain long-term, covert access to networks[4]. At the same time, AI adoption is accelerating, with 68% of large companies using at least one AI technology in 2023[5].

The UK's NCSC highlights that AI systems introduce a significantly larger attack surface than traditional software, due to their complexity, data reliance, and rapid development[6]. These systems are vulnerable to a range of adversarial attacks that can manipulate behaviour, extract sensitive data ,or degrade performance, with impacts that can cascade across interconnected systems.

These risks, including data leakage, insecure code generation, and adversarial manipulation, may become embedded at scale as AI-enabled systems are adopted[7]. Our primary research also highlights a growing reliance on AI-generated code, which may introduce new quality and security risks into critical environments.

While traditional cyber security controls remain essential, AI introduces unique vulnerabilities that require dedicated and still-maturing security approaches. NCSC guidance emphasises the need to treat AI models, data and supporting systems as high-value assets, requiring strong security controls and continuous monitoring to build long-term resilience[8].

SECURING AI ADOPTION ACROSS IT & OT

Digital Twins (dynamic virtual replicas of environments) enhance visualisation for CNI companies and are employed by NATO to simulate cyberattack scenarios[9]. These virtual

models enable scenario rehearsal, virtual update validation, and coordinated recovery drills. Similarly, Internet of Things (IoT) and edge devices improve responsiveness by bringing decision making closer to the application. Yet, these technologies can also broaden the attack surface by introducing valuable assets and sensitive training data that require protection. CNI operators can preserve the benefits of these AI-driven technologies while mitigating misuse with strong governance, segmented access, and where feasible, synthetic data.

As CNI operators integrate IT and OT to drive efficiency, such as remotely monitoring and managing field equipment, they expand the pathways between enterprise networks, industrial control systems, and their connecting data flows. AI has mostly been built and secured in IT systems, where security is optimised for confidentiality, integrity, and availability[10]. When AI works within OT, the main priorities shift to the latter, emphasising safety and keeping systems running reliably.

OT systems are often older, less secure, and harder to integrate with modern AI tools. This creates a practical friction point: moving OT data into AI services often requires bridging modern IT cryptography and key management with OT environments that may rely on legacy protocols, constrained devices, and long-lived systems. If not engineered carefully, this can weaken transport protections or force insecure gateways. Therefore, treating IT and OT as a single uniform "infrastructure" when adopting AI increases the likelihood of cascading failures, misclassification of sensitive operational data, and unintended access or control. Security controls that are explicitly designed for OT connectivity and AI-driven workflows are essential at every stage of design and deployment.

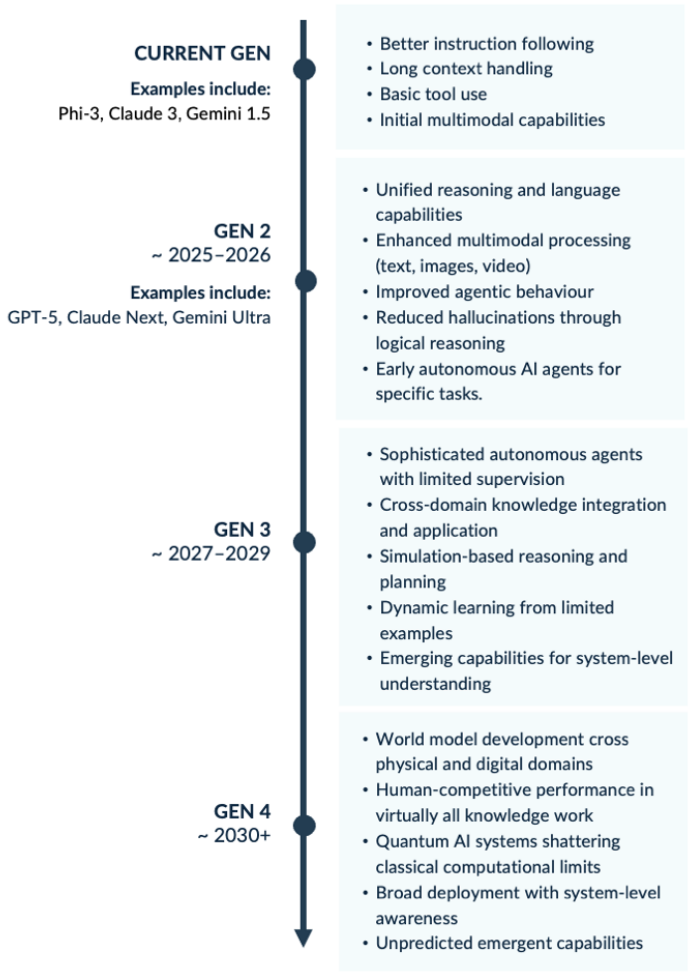
AI IN PRACTICE: RISING ADOPTION AND EVOLVING RISK

Research is being conducted to address the secure deployment of AI, specifically generative AI, as it continues to grow and become embedded in critical systems. In 2023, Laura Weidinger and colleagues at Google DeepMind led a research effort to categorise AI security evaluations for generative AI systems, finding that most evaluations touch on a wide range of potential risk scenarios, but in limited detail, highlighting both the scale and complexity of the challenge ahead[11].

2024 saw the rise of large language models (LLMs) and chatbots, which evolved from text-based tools to multimodal systems capable of understanding and generating text, images and video. In 2025, attention turned to agentic AI, as organisations explored

automation and decision-making through intelligent, autonomous models. Interest is also growing in reinforcement learning, where agents learn through interaction rather than labelled data. Following the rise of agentic AI, the development and application of multi-agent systems can be seen throughout the market in 2026, where workflows often incorporate interconnected agents for more sophisticated use cases.

This can be seen through the launch of the OWASP Agentic Research Council at Infosec in June 2026, where the cross-cutting challenges of multi-agent systems are a focal point of discussion and development[12], alongside global governance frameworks like the NIST AI RMF and ISO/IEC 4200[13], which are driving standardised approaches to secure deployment, risk management, and accountable AI operations.



Building on LLMs, some companies are developing small language models - these are compact versions designed for operational technology with limited bandwidth. Looking ahead, research into the convergence of AI and computing promises to open new frontiers in capability and performance.

Figure 1: Timeline of evolving AI technologies from current to medium-long term.

ADVERSARIAL USE OF AI: A LOWERED BARRIER TO ENTRY

AI is also lowering the barrier to entry for attackers. Adversaries are increasingly using AI to automate reconnaissance, generate targeted attacks, and accelerate exploitation.

This is lowering barriers to entry and amplifying the scale of attack[14], particularly in social engineering and vulnerability discovery. AI-driven phishing campaigns are significantly more effective and profitable than traditional methods [15], while the growing use of automation is accelerating the speed and reach of cyberattacks across complex infrastructure systems[16].

Threat actors are leveraging AI as a malware development accelerator, supporting iterative engineering tasks across the malware lifecycle. Emerging technologies such as agentic AI are enabling persistent, adaptive and hard-to-detect threats. Real-world examples and proof-of-concept experiments illustrate the potential for these systems to support automated reconnaissance, infrastructure management, malware development, and post-compromise decision-making[17]. Recent developments around Anthropic's "Mythos" model autonomously identifying and exploiting previously unknown software vulnerabilities illustrate the advancing capability that could be utilised by attackers, reinforcing the trend of a lowered barrier to entry for sophisticated incidents.[18] Cyberattack examples such as Jaguar Land Rover[19], Marks and Spencer[20], Co-op[21] and Harrods[22], while not explicitly due to AI misuse, are a reminder of the operational and business impact from cyberattacks such as ransomware.

AI FOR CYBER DEFENCE: AN OPPORTUNITY TO EMPOWER DEFENDERS

While AI is a powerful tool for attackers, it is also a gamechanger for cyber security defence. Organisations are using AI to reduce the time to detect, respond to, and recover from attacks, helping them keep pace with more sophisticated adversaries. Emerging research highlights the potential of autonomous cyber defence, using techniques such as deep reinforcement learning and large language models to identify vulnerabilities and automate detection and response[23].

Government and standards bodies increasingly recognise this dual role, framing AI as both a source of risk and a defensive opportunity [24] [25]. For CNI operators, the opportunity is not simply to automate existing security processes, but to use AI to improve visibility across fragmented, high-volume, and time-sensitive environments. AI-enabled tools can support continuous monitoring, correlate weak signals across IT and OT systems, prioritise alerts, and help analysts identify abnormal behaviour before it escalates into operational disruption.

This is particularly important where defenders face growing volumes of telemetry, legacy infrastructure constraints, and faster adversarial activity. Used responsibly, AI

can strengthen defensive posture by augmenting human analysts, accelerating triage, and enabling more adaptive response across complex infrastructure networks.

Taken together, AI in CNI represents a dual-use capability. While it expands the attack surface and introduces new vulnerabilities, it also strengthens defensive capability. The challenge for operators is to adopt AI securely and strategically, in line with guidance[26] [27], ensuring it reduces risk rather than amplifying systemic fragility across critical infrastructure.

COMBINED RISKS TO CNI FROM AI

Understanding how AI reshapes the cyber threat landscape requires a framework that captures both the evolving tactics of adversaries and the vulnerabilities introduced through integration.

Traditional models, such as the Cyber Kill Chain[28], the World Economic Forum's cyber risk frameworks,[29] NIST's Cybersecurity Framework Profile for Artificial Intelligence (Cyber AI Profile)[30] and regulatory baselines like the European Union's NIS2 Directive[31] have provided a foundation for understanding attacker behaviour and organisational vulnerabilities, but they are limited in explaining the complexity of modern, multi-vector operations.

The UK's NCSC introduced a common taxonomy for understanding adversarial attacks against machine learning and AI[32]. It highlights that AI systems create a significantly larger attack surface than traditional software due to their complexity, data reliance, and rapid development.

It identifies a range of adversarial attack types that can manipulate models, extract sensitive data, or alter system behaviour, potentially cascading across interconnected systems. While traditional cyber security controls remain essential, AI introduces unique vulnerabilities that require dedicated, still-maturing security approaches across the system lifecycle (as discussed in ETSI EN 304 223[33]).

Applying this framework to specific networks can help defenders anticipate adversary tactics, improve detection, and strengthen response strategies. This framework has been used to explore how AI both augments adversarial threats, creates novel risks within CNI environments, and enables defensive responses.

Particularly important insights emerge when mapping risks against the eight malicious actor goals defined in the framework. Unlike earlier models that take a technology-agnostic view of attack pathways, this approach focuses on AI-specific risks and behaviours, such as "attribution of output," where attackers attempt to identify model-generated outputs, through hidden watermarks for example.

In reconnaissance, AI compresses attacker timelines by enabling rapid processing of large datasets and complex networks. Increased AI adoption within CNI environments may also create clearer attack footprints. At the same time, defenders can leverage AI for anomaly detection, containment, and incident response. This duality underscores the importance of early-phase resilience, where strong cyber security hygiene remains critical.

AI is increasingly transforming the early stages of cyberattacks, enabling threat actors to waste resources and degrade performance of the victim system with greater speed, precision, and accessibility. It enhances evasion techniques and the embedding of hidden behaviours, including polymorphic malware. While self-learning malware improves persistence, AI-driven defences are narrowing the gap in detection and response.

Overall, AI is likely to democratise attack capabilities, reducing the gap between less experienced and advanced actors. This lowers the barrier to conducting impactful cyberattacks against CNI, extending disruptive capabilities not only to individuals but also to states that previously lacked such capacity. This points to a broader escalation in both individual and state-led cyber activity, raising the risk of wider geopolitical consequences.

Therefore, AI augments both offensive and defensive operations. The net impact depends on deployment choices, governance, and operational maturity, with benefits accruing to either side depending on use.

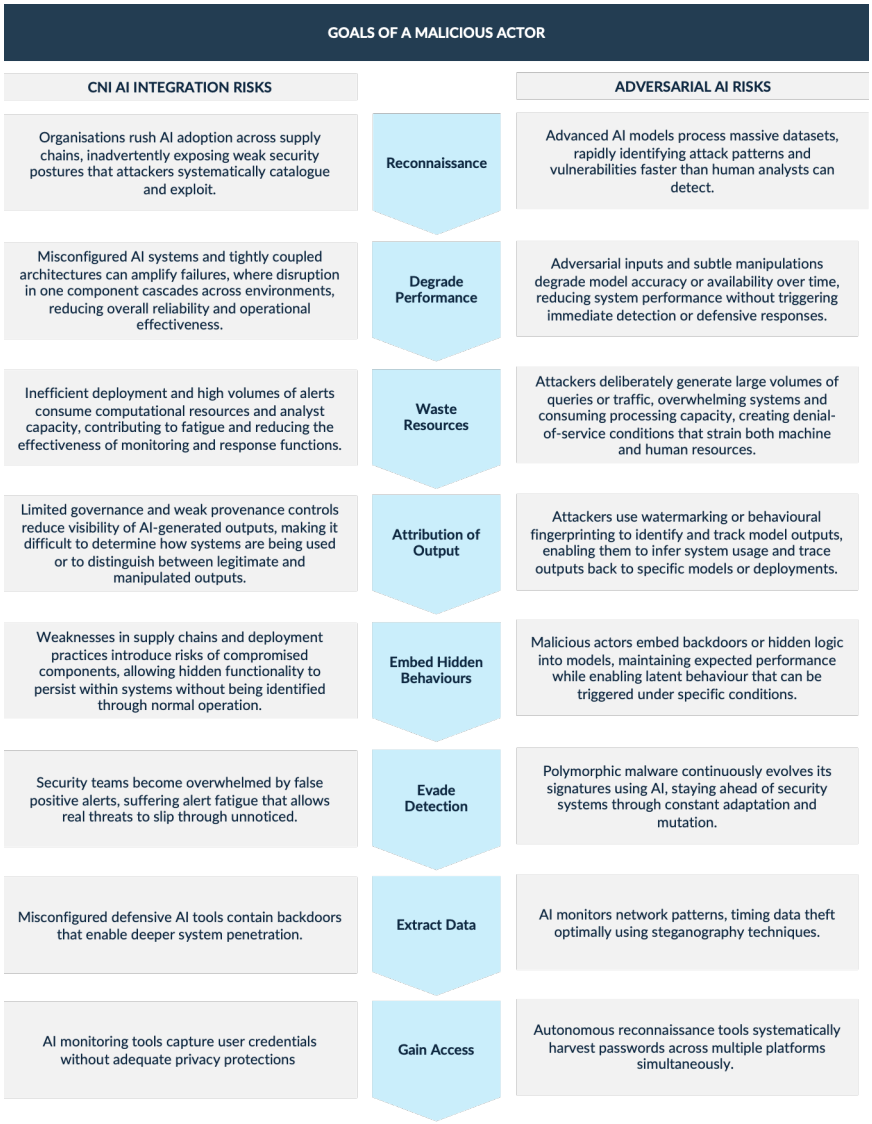
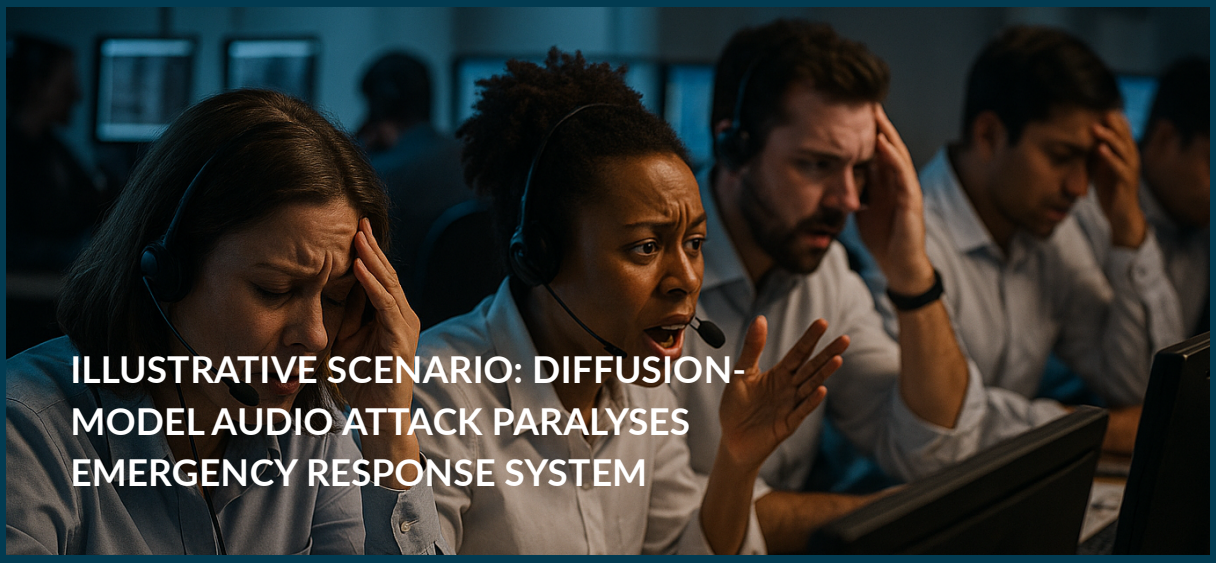


Figure 2: Adoption and adversarial AI risks mapping along the 8 goals of a malicious actor as outlined by the UK NCSC.



ILLUSTRATIVE SCENARIO: DIFFUSION-MODEL AUDIO ATTACK PARALYSES EMERGENCY RESPONSE SYSTEM

Emergency services were temporarily paralysed by a coordinated flood of AI-generated voice calls, mimicking urgent distress scenarios. The attack was launched by EchoShade, a non-state-backed, decentralised hacker collective known for repurposing open-source AI tools. It was classified as the UK's first diffusion-model audio attack on CNI, highlighting how non-state actors can now weaponise advanced generative AI, widening the threat landscape beyond traditional cyber operations.

ILLUSTRATIVE SCENARIO: CRYBERCRIME GROUP MONETISES FORGOTTEN BREACH DATA FROM ENERGY AND WATER SECTORS

A cybercrime group known as ReZero is profiting from legacy breach data stolen years ago from energy, water and transport operators across Europe and North America. Rather than hacking new systems, the group uses AI to exploit "digital leftovers" such as unrevoked credentials, unchanged administrator passwords, system diagrams ,and outdated but still valid engineering documents. Using AI, ReZero indexes and repackages this material for sale on darknet markets, enabling follow-on ransomware and industrial control system attacks. Investigators have linked recent intrusions, including an attack on a regional water utility, to this resold data. The case highlights that historical breach data remains an active security risk, not a closed incident.

CNI SECTORS AT RISK

SECTORAL BREAKDOWN: OUTLINING KEY CRITERIA FOR RISK ANALYSIS

While adversarial techniques are consistent across domains, their impact is shaped by sector-specific characteristics, including reliance on operational technology (OT), system interdependencies, and the pace of AI adoption.

Each sector faces distinct contextual risks, be that managing high-value, time-sensitive assets in the financial sector or navigating OT dependencies in telecoms and transport. We sought to prioritise sectors of high strategic importance and interdependence.

The following criteria were used to identify which sector was most at risk of adversarial use of AI or risks arising from adopted AI solutions:

- **High value targets:** sectors of strategic importance that are essential for societal functioning, including a high reliance on other sectors.
- **Legacy infrastructure & digitisation gap:** sectors that have a high volume of legacy digital infrastructure that is still in operational use or immature in digital transformation activities, including digitising their OT.
- **Complex supply chains:** sectors that have a high reliance on networks of suppliers and contractors. AI can identify and exploit weak links (e.g., through phishing, social engineering, or software supply chain attacks).
- **Lack of regulation and standards:** many sectors lack clear AI-specific cyber security policies, especially around explainability, access control, and adversarial robustness.
- **Insider threats enhanced by AI:** sectors that rely heavily on humans-in-the-loop, where AI can help adversaries coerce or co-opt insiders through psychological profiling or behavioural manipulation or malicious insiders can use AI with bad intent.
- **Psychological and societal impact:** sectors at risk of causing diminished public trust or societal damage, such as disinformation campaigns that undermine trust (e.g., false alerts about water contamination or power grid failures).

Energy has prevalent risks across all criteria. Water and Healthcare are also particularly at risk due to the immaturity of the sector's digitisation. Transport, Telecoms and Finance are at risk primarily through their extensive supply chains and impact on other sectors.

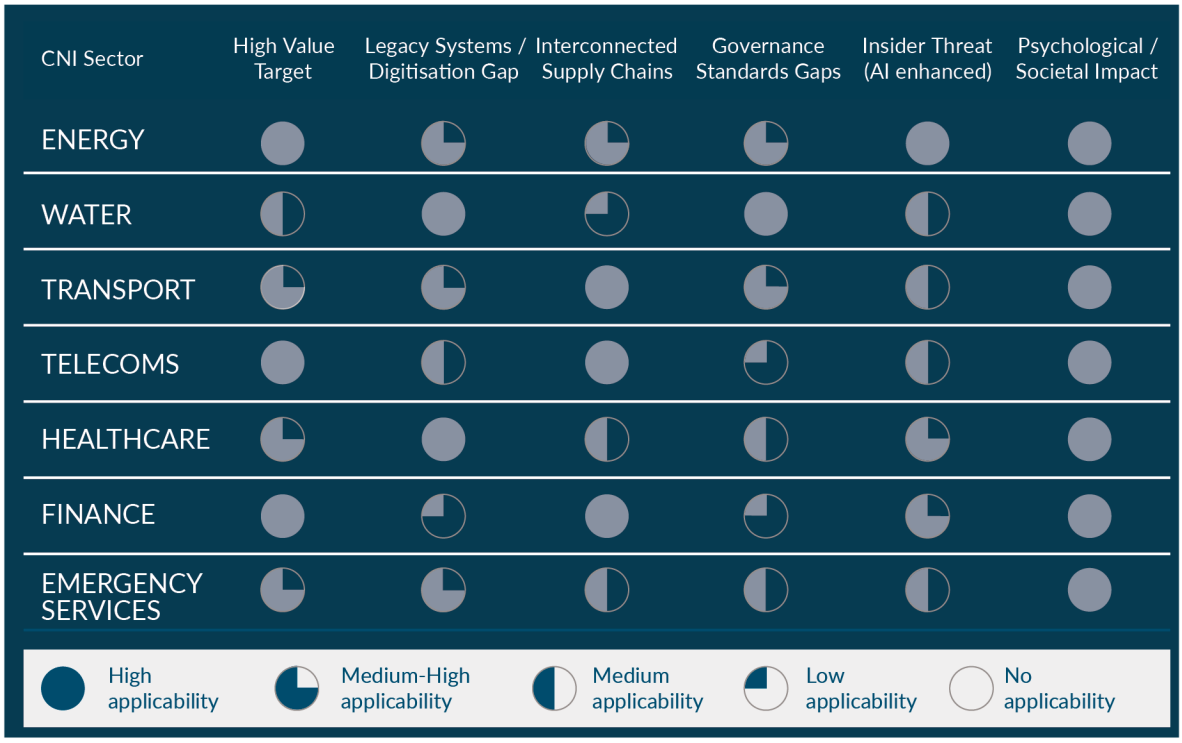


Figure 3: High priority CNI sectors scored against risk criteria.

RISKS BY CNI SECTORS: HOW AI RISK VARIES ACROSS INFRASTRUCTURE SECTORS

Table 1: Comparative overview of AI-related risks, drivers and vulnerabilities across CNI sectors.

	RISK LEVEL	RISK DRIVERS	KEY VULNERABILITIES
ENERGY	**Systemic and Highest Impact Risk** The energy sector represents a highly critical systemic risk. Its combination of high consequence, complexity and increasing automation makes it especially vulnerable to large-scale, cascading failures.	• High reliance from other systems • Extensive use of complex OT environments • Rapid deployment of AI in grid management, balancing and trading • Increasing use of distributed and edge AI systems	• Energy balancing platforms as high-value targets for manipulation • Model drift and subtle "micro-nudges" that alter system behaviour over time • Human vulnerabilities in control centres, particularly through phishing and deepfakes • Legacy infrastructure and delayed remediation of known vulnerabilities • Risks from poorly trained or insufficiently tested AI models causing instability or cascading failures
HEALTHCARE	**High Vulnerability, Low Maturity** Healthcare systems are characterised by widespread legacy infrastructure and rapid, often uncoordinated, AI adoption. Vulnerabilities are amplified by weak cyber hygiene and fragmented governance approaches.	• Extensive use of legacy systems connected to networks • Growing adoption of AI-enabled medical devices and digital health tools • Expansion of third-party AI integrations and supply chain dependencies • Limited visibility and governance over bespoke AI tools developed by clinicians	• Unpatched systems lacking basic protections such as MFA • Large internet-connected attack surface across medical IoT devices • Emergence of "shadow AI" tools developed outside formal governance • Third-party AI integrations introducing supply chain vulnerabilities • Legacy systems that may be incompatible with modern AI security controls
WATER	**High Vulnerability, Fragmented Risk Landscape** The water sector faces significant risk driven by fragmentation, inconsistent cyber maturity and uneven AI adoption across operators. Structural inconsistency creates gaps that adversaries may exploit.	• Security maturity varies widely across operators • Legacy infrastructure complicates secure AI integration • Inconsistent cyber investment, skills and IT/OT architectures • Limited regulatory coverage across parts of the sector	• Sensor manipulation and AI-enabled alert flooding, contributing to alert fatigue • Exploitation of operational data such as reservoir status information • Legacy systems lacking resilience when interfaced with AI-enabled tools • Limited coordination and inconsistent standards across operators • AI-driven disruptions that may go undetected due to fragmented oversight

FINANCE	**Advanced but Complex** Financial services is among the most mature sectors in AI adoption, but this maturity introduces significant complexity and interconnected risk across markets, trading systems and customer operations.	• Extensive AI deployment across trading, fraud detection and customer services • High-speed, AI-driven market activity operating beyond human timescales • Reliance on large volumes of proprietary AI models • Strong interdependencies with payment and financial market infrastructure	• Deepfake-enabled fraud and impersonation • AI-driven market manipulation and disinformation campaigns • Algorithmic instability and unintended interactions between trading systems • Risks of inadvertent algorithmic collusion and emergent behaviour • AI training data lakes and proprietary models as high-value targets
TELECOMS	**System and Supply Chain Risk** The communications sector is defined by extensive interdependencies, legacy infrastructure and increasing reliance on AI-enabled operational systems. Disruption could create cascading impacts across multiple sectors.	• Continued reliance on legacy telecom infrastructure • Growing use of AI for predictive fault detection and network optimisation • AI embedded within operational network management systems • Dependence on complex supply chains and backbone infrastructure	• Exploitation of outdated and poorly patched infrastructure • Manipulation of AI systems that guide engineers or automate network responses • False alerts or suppression of genuine faults within AI-enabled monitoring systems • Reduced visibility and oversight over embedded AI decision-making • Nation-state targeting of critical routing and backbone infrastructure
EMERGENCY SERVICES	**Emerging Risk Through Transition** Emergency services currently have lower levels of AI integration, but risks are expected to increase as the sector modernises and transitions towards new communications infrastructure.	• Ongoing transition from legacy systems to commercial 4G-based infrastructure • Growing interest in future AI-enabled decision-support capabilities • Dependence on voice-based communications vulnerable to synthetic media attacks	• AI-generated voice spoofing and deepfake emergency calls • Increased exposure associated with migration to commercial network infrastructure • Legacy Airwave system fragility and ageing infrastructure • Future risks from AI-enabled decision-support systems if cyber security is not prioritised during deployment

THE IMPACT OF OT ARCHITECTURES

While there are many emerging cross-sector risks from adversarial use of AI and AI integration into CNI sectors, some risks are specific to a sector due to factors such as their position within a wider ecosystem and other sectors' reliance on it, their dependency on operational technology, and the pace of AI adoption within that sector.

As sectors with higher OT usage were found to have more security gaps, a deep dive into this architecture was conducted to outline areas for further attention.

The Purdue model[34] shows the layers that make up OT. These layers are connected; however, they require very different protocols and security. OT devices are typically user-unfriendly and require different skillsets and software.

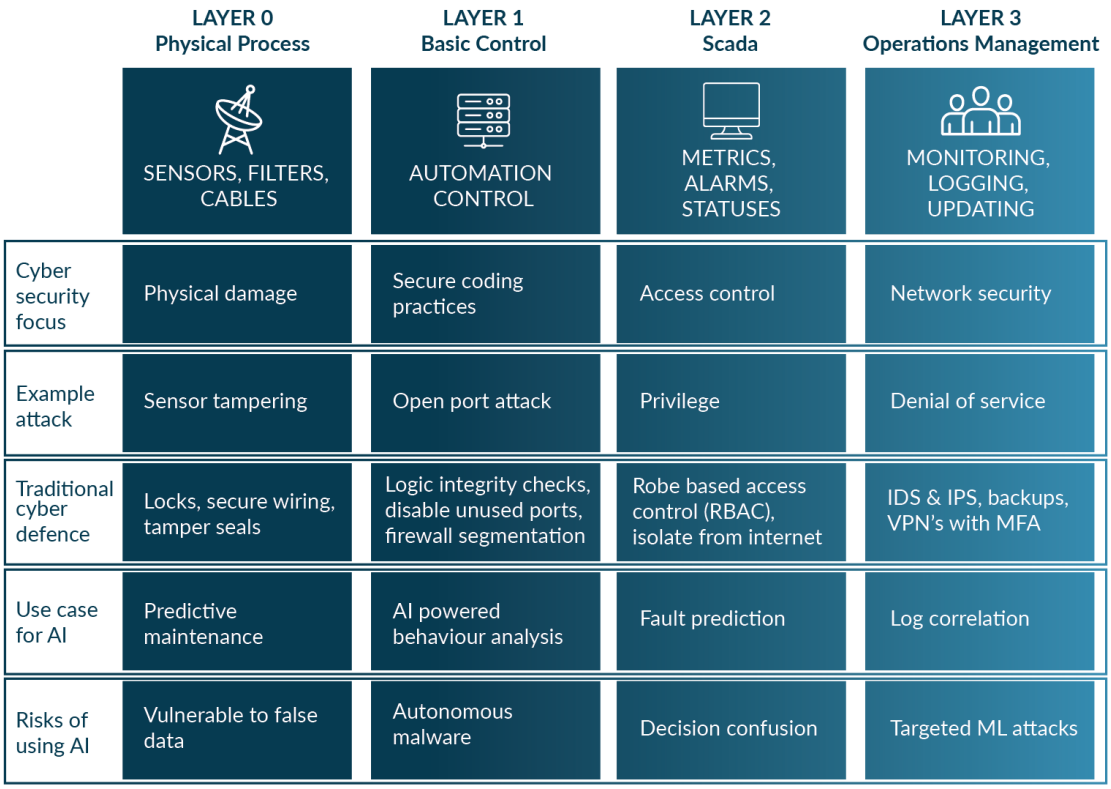


Figure 4: The first three layers of the Purdue Model mapped against key cyber security themes.

While AI has so far been concentrated in IT environments, rapid adoption is now driving its deployment deeper into OT layers, from manufacturing to energy. This integration requires AI models to access and process large amounts of operational data from sensors and networks across both IT and OT domains. Such access raises fundamental questions of data governance, including what data is shared, who controls it, and where it is stored and processed. Poor data governance increases the risk of inadvertently exposing CNI data to unauthorised or hostile actors.

RECOMMENDATIONS

The following strategic recommendations have been identified as relevant across sectors, operators, and solution providers:

1. **Embed security by design.** Integrate robust security and governance such as ETSI EN 304223 from the outset of AI development, ensuring resilience against cascading failures in CNI.
2. **Improve AI security and operational constraints.** Explore product innovations that could strengthen security and operational constraints within AI systems, potentially supported by hardware-based security measures where appropriate, to reduce the risk of misuse in critical environments without relying on static hardcoded limits.
3. **Modernise legacy infrastructure**. Bridge the gap between AI and outdated OT systems by providing intelligent interoperability layers, predictive insights, enhanced oversight and secure, operator-controlled handoff mechanisms that enable safe integration of advanced capabilities without requiring wholesale replacement of ageing systems.
4. **Adopt risk-based release frameworks**. Roll out AI in phased, transparent stages, prioritising cyber secure risk management over speed when integrating into safety-critical systems.
5. **Adopt risk-based security outcomes.** Support measured, risk-based security requirements prior to AI deployment to promote consistent practice and reduce the likelihood of later security challenges.
6. **Deploy AI to strengthen cyber defence**. Work on the development and deployment of secure and robust AI systems to enhance vulnerability detection, anomaly monitoring, and rapid incident response at machine speed.
7. **Prioritise security, transparency and responsible use in AI diffusion**. Explore replicating the successful cloud 'shared responsibility model' for AI systems. A shared responsibility model for AI should mirror the clarity and collaboration that transformed cloud security, but with an emphasis on transparency and meaningful customer choice. Providing customers with actionable information on risks, safeguards and configuration options across the AI lifecycle.
8. **Deliver targeted AI training**. Fund and embed role-specific training in high-risk sectors to build operational competence, reducing misuse, and over-reliance on AI systems.

9. **Collaborate on secure procurement standards**. Advance standards that are harmonised, that can guide risk-based outcomes and encourage secure release practices through procurement levers and voluntary best practices. This can support raising the floor without stifling innovation.

10. **Support the development of UK AI resiliency**. Support transparency when activating AI functions, reducing risks from hidden or unassessed integrations.

CONCLUSION

Although AI has the potential to introduce novel risks, it also represents a transformational opportunity for CNI organisations to strengthen resilience, modernise operations, and improve cyber defence at scale and speed.

When deployed with proactive governance, security-by-design principles, and robust oversight, AI can provide several benefits: 1. Enhance situational awareness across complex environments; 2. Provide predictive insights that extend the life and security of legacy infrastructure; 3. Automate the detection of anomalies and emerging threats; and 4. Enable a faster, more coordinated, incident response.

It can also support continuous monitoring at scale, reduce analyst burden through intelligent automation, improve threat intelligence correlation across fragmented data sources, and enable more adaptive and self-healing systems that respond dynamically to evolving risks.

Combined with risk-based release frameworks, secure interoperability mechanisms and collaboration with trusted AI providers, these capabilities allow CNI operators not only to counter accelerating adversarial activity, but also to build more reliable, efficient and future-proof critical services in the UK and beyond.

APPENDIX

To ensure diverse perspectives on operational risks, policy requirements, and innovation pipelines, a mixed-methods research approach was used combining primary research, stakeholder engagement and targeted literature review to capture the complexity of emerging AI risks.

Plexal interviewed 30 experts across CNI operators (9), government (7), academic researchers (5), startup/SME participants (6), cross-sector experts (2) and AI Suppliers (1). Plexal also surveyed a further 15 CNI experts and carried out two structured workshops that supplemented these insights: an initial His Majesty's Government (HMG) and academic convening in March 2025 with 15 participants establishing foundational frameworks, and a CNI stakeholder workshop in May 2025 bringing together over 30 cross-sector participants to validate findings and generate sector-specific risk assessments.

TIER	SECTORS	RATIONALE
Tier 1 **(Highest priority)**	Energy, Finance, Health, Transport, Water, Communications	Sectors of strategic importance which are essential for societal functioning including high reliance from other sectors
Tier 2 **(Medium priority)**	Emergency Services, Government, IT	High societal impact, digitisation gaps and high visibility and trust
Tier 3 **(Lower priority)**	Food, Defence, Space, Chemicals, Civil Nuclear, Commercial, Manufacturing, Dams	Important sectors, but less at risk due to factors such as high regulations or less integral into wider societal effects

Table 2: CNI sectors prioritised for this research.

REFERENCES

[1] House of Commons Library, Artificial Intelligence in government departments, 2025.

[2] Understanding adversarial attacks against Machine Learning and AI, NCSC, 2026

[3] Cybersecurity & Infrastructure Security Agency, Countering Chinese State-Sponsored Actors Compromise of Networks Worldwide to Feed Global Espionage System, 2025.

[4] Anthropic, Disrupting the first reported AI-orchestrated cyber espionage campaign, 2025

[5] Forbes, UK Artificial Intelligence (AI) Statistics and Trends in 2025, 2025.

[6] Understanding adversarial attacks against Machine Learning and AI, NCSC, 2026

[7] Abdali et al., "Securing Large Language Models: Threats, Vulnerabilities and Responsible Practices", arxiv, 2024.

[8] HM Government, AI Cyber Security Code of Practice, 2025.

[9] Cooperative Cyber Defence Centre of Excellence: Locked Shields Exercise Showcased the Power of Cooperative Defence, 2024

[10] Fortinet, What Is the CIA Triad and Why Is It Important?

[11] Griffin, C., "What We Learned from Reading ~100 AI Safety Evaluations", Google DeepMind AI Policy Perspectives (2025).

[12] Infosecurity Europe, OWASP Gen AI Security Summit, 2026

[13] ISO/IEC 42001 vs. NIST AI RMF: A Comparative Analysis

[14] AI as tradecraft: How threat actors operationalize AI | Microsoft Security Blog

[15] Microsoft Digital Defense Report | Microsoft

[16] ENISA Threat Landscape 2025

[17] Ibid. See also: Disrupting the first reported AI-orchestrated cyber espionage campaign \ Anthropic

[18] McMahon, L. and Tidy, J., What is Claude Mythos and what risks does it pose?, BBC News 2026

[19] Jaguar Land Rover posts heavy loss after cyber-attack | BBC News Article

[20] M&S profits almost wiped out after cyber hack hit sales | BBC News Article

[21] Co-op says cyber-attack cost it £206m in lost sales | BBC News Article

[22] Hackers contact Harrods after 430,000 customer records hit by IT breach | BBC News Article

[23] AI for Cyber Defence Research Centre - AICD

[24] President Trump's Cyber Strategy for America

[25] The near-term impact of AI on the cyber threat | NCSC

[26] Guidelines for secure AI development | NCSC Guidance

[27] Code of Practice for the Cyber Security of AI | UK Government Guidance

[28] Microsoft, "What is the cyber kill chain?", accessed September 2025.

[29] World Economic Forum, Global Cybersecurity Outlook 2025, 2025.

[30] IR 8596, Cybersecurity Framework Profile for Artificial Intelligence (Cyber AI Profile): NIST Community Profile | CSRC

[31] The NIS 2 Directive, accessed September 2025.

[32] UK NCSC, Understanding Adversarial Attacks Against Machine Learning and AI

[33] European Telecommunications Standards Institute (ETSI), Securing Artificial Intelligence (SAI): Baseline Cyber Security Requirements for AI Models and Systems, ETSI EN 304 223 V2.1.1, December 2025

[34] Williams, T. J., The Purdue Enterprise Reference Architecture (PERA), 1992.