DIGITAL POLICY ALLIANCE

www.DigitalPolicyAlliance.org

ROBUST AI:

A FRAMEWORK FOR UK LEADERS AND POLICY MAKERS

AI ESSENTIALS WORKSHOP – APRIL 2026

TABLE OF CONTENTS

INTRODUCTION: WHY ROBUST AI MATTERS NOW

As AI systems move from experimental pilots to mission-critical infrastructure, their robustness becomes a matter of operational survival. When a customer service chatbot hallucinates, the consequences are manageable. When an AI system managing hospital patient flow fails, or a financial trading algorithm behaves unpredictably, or an autonomous vehicle misinterprets sensor data, the stakes are existential for organisations and potentially life-threatening for individuals [1,2,3].

The UK's increasing reliance on AI across critical sectors creates new categories of risk. Recent incidents illustrate the challenge: AI-powered fraud detection systems at major UK banks have experienced periods of degraded performance during peak transaction times [4]; NHS trusts deploying AI diagnostic tools have reported inconsistent results across different patient populations [5]; and automated decision systems in local government have produced unexpected outputs when encountering edge cases not represented in training data [6]. These aren't hypothetical scenarios. They represent the current reality of AI deployment at scale.

Robust AI is about deliberately managing four interconnected technical challenges: **security vulnerabilities** that adversaries can exploit; **reliability failures** that degrade performance unpredictably; **data quality issues** that undermine system accuracy; and **operational fragility** that prevents graceful degradation under stress. Addressing these challenges requires engineering discipline, governance frameworks, and procurement practices that prioritise resilience alongside capability.

The 4 Pillars of Robust AI

The framework described here draws on evidence from UK incidents, international standards including ISO/IEC 42001, and research from the National Cyber Security Centre, Alan Turing Institute, and leading industry practitioners. It provides senior leaders and policy makers with the technical grounding needed to ask the right questions and make informed decisions about AI robustness.

TECHNICAL FOUNDATION OF AI ROBUSTNESS

Before examining specific robustness dimensions, it is essential to understand why AI systems are inherently more fragile than traditional software, and what technical characteristics create vulnerability.

Why AI Systems are Different

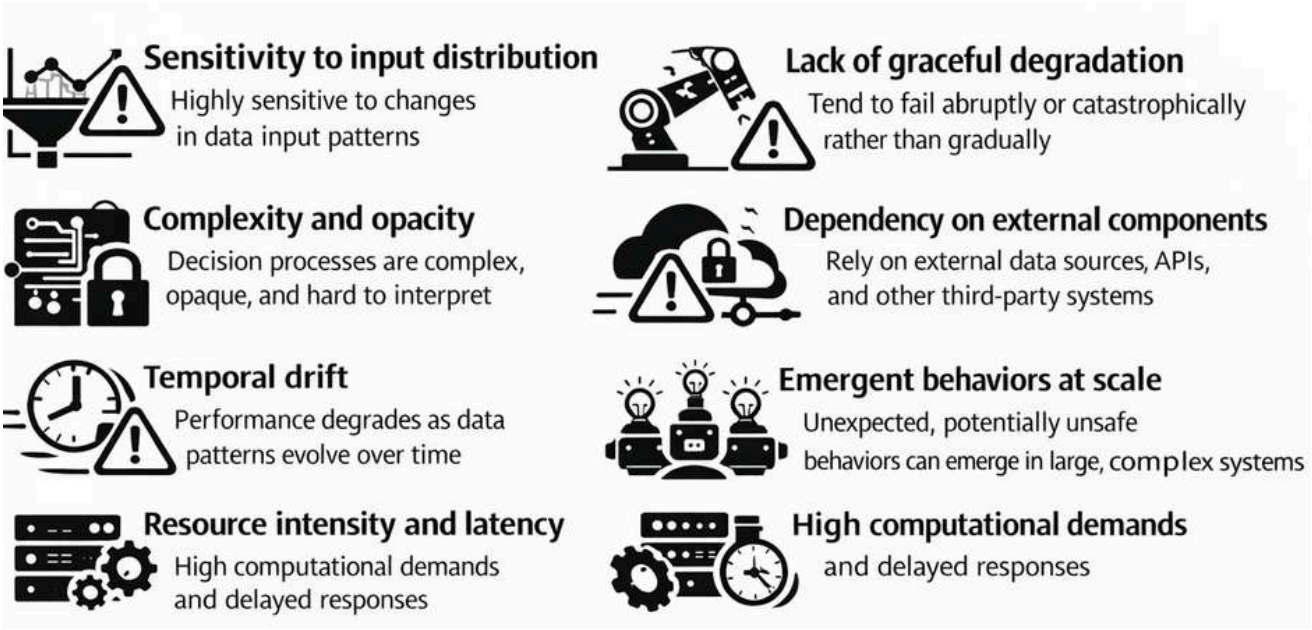
Traditional software follows explicit, deterministic rules written by programmers. When something goes wrong, engineers can trace the logic, identify the bug, and fix it. AI systems, particularly those based on machine learning and large language models, operate differently. They learn patterns from data rather than following explicit rules, making their behaviour harder to predict and debug [7].

This fundamental difference has profound implications for robustness. A traditional software bug typically produces the same wrong output every time given the same input. An AI system might produce different outputs for similar inputs, might fail in ways that depend on subtle characteristics of the data, and might degrade gradually rather than failing continuously.

Furthermore, AI systems are statistical in nature. They don't "know" facts the way humans do; they have learned correlations and patterns that usually produce appropriate outputs. This means they can fail in ways that seem inexplicable, producing confident-sounding nonsense, missing obvious patterns, or behaving unpredictably when encountering situations outside their training profile.

The Attack Surface Problem

Additionally, AI systems present a dramatically expanded attack surface compared to traditional software. Beyond conventional cybersecurity vulnerabilities (network intrusion, authentication bypass, SQL injection, etc.), AI systems are vulnerable to attacks that exploit their learning mechanisms [8]. Adversaries can craft inputs specifically designed to cause misclassification, manipulate training data to poison model behaviour, or extract sensitive information embedded in model parameters.



The Foundations of Robust AI

Key Technical Characteristics That Affect Robustness

Understanding these core aspects of AI matters for robustness because modern AI systems have characteristics that create specific vulnerabilities.

Sensitivity to input distribution: AI systems perform well on data similar to their training data but can fail unpredictably on out-of-distribution inputs. A facial recognition system trained on well-lit, frontal photographs may fail dramatically on images taken in different lighting conditions or angles. This isn't a bug that can be simply fixed; it's inherent to how these systems learn [9].

Lack of graceful degradation: Traditional systems often fail in predictable ways that allow for error handling. AI systems can fail silently, producing outputs that look plausible but are completely wrong. For example, a language model might generate confident-sounding medical advice that is factually incorrect, with no indication to the user that something has gone wrong [10].

Complexity and opacity: Modern AI models contain billions of parameters, making them impossible to fully audit or understand. Even their creators cannot always explain why they made specific decisions. This opacity creates challenges for security (attackers can exploit unknown vulnerabilities), reliability (failures may occur in unpredictable circumstances), and debugging (root cause analysis becomes extremely difficult) [11].

Dependency on external components: AI systems typically depend on complex software stacks, cloud infrastructure, and third-party services. A vulnerability in any component can compromise the entire system. The supply chain for AI includes pre-trained models, datasets, and tools that organisations may not fully control or understand [12].

Temporal drift: AI systems can degrade over time as the real-world data they encounter diverges from their training data. A fraud detection model trained on 2023 transaction patterns may perform poorly against 2026 fraud techniques. Without continuous monitoring and retraining, performance erosion is inevitable [13].

Emergent behaviours at scale: Large AI systems can exhibit behaviours that were not present in smaller versions or during testing. Capabilities (and vulnerabilities) can emerge unpredictably as models grow larger, making it difficult to anticipate how systems will behave in production [14].

Resource intensity and latency: AI inference requires substantial computational resources, creating potential bottlenecks under load. Systems designed for average traffic may fail during peak demand, precisely when reliable performance is most critical. Latency variations can also cause cascading failures in systems that depend on timely AI outputs [15].

Why This Technical Foundation Matters

These characteristics highlight four key areas for building robust AI systems:

- **Security and adversarial resilience challenges** arise because AI systems can be attacked through their learning mechanisms, not just cyber vulnerabilities.
- **Reliability and performance problems** stem from the statistical nature of AI and its sensitivity to data distribution, creating unpredictable failure modes.
- **Data integrity and quality issues** arise because AI systems are fundamentally dependent on training data, and poor data produces unreliable outputs.
- **Operational resilience requirements** increase because AI systems introduce new failure modes that require new monitoring, incident response, and recovery.

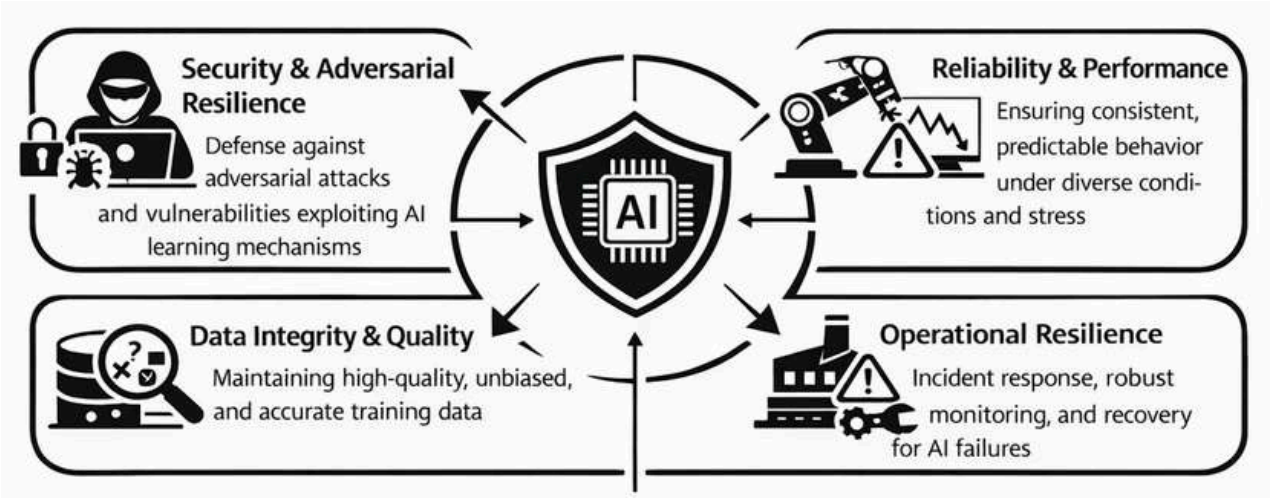
None of these challenges can be addressed purely through technical fixes. All require governance frameworks, procurement standards, and operational practices that embed robustness considerations throughout the AI lifecycle.

Good Practice Reduces Risk — It Does Not Remove It

The approaches in this framework are worth adopting and organisations that follow them will be in a stronger position than those that don't. But none of these measures guarantees safety. AI systems can still fail in unexpected ways, even when best practice has been followed, and some failures only emerge once a system is operational.

This has two practical implications for decision-makers:

- **Before deployment:** Ask not just "have we followed good practice?" but "what could still go wrong, and is that risk acceptable?". In some cases, the honest answer will be that deployment should not proceed.
- **After deployment:** Treat ongoing monitoring as a permanent commitment, not a one-time checkbox.



Building Robust AI Systems

SECURITY AND ADVERSARIAL RESILIENCE: PROTECTING AGAINST THREATS

The Challenge

AI systems face a unique threat landscape that combines traditional cybersecurity risks with novel attack approaches specific to machine learning. Adversarial attacks—carefully crafted inputs designed to cause AI systems to fail—have moved from academic research to practical concern. Nation-state actors and sophisticated cybercriminals are actively developing capabilities to exploit AI vulnerabilities [16].

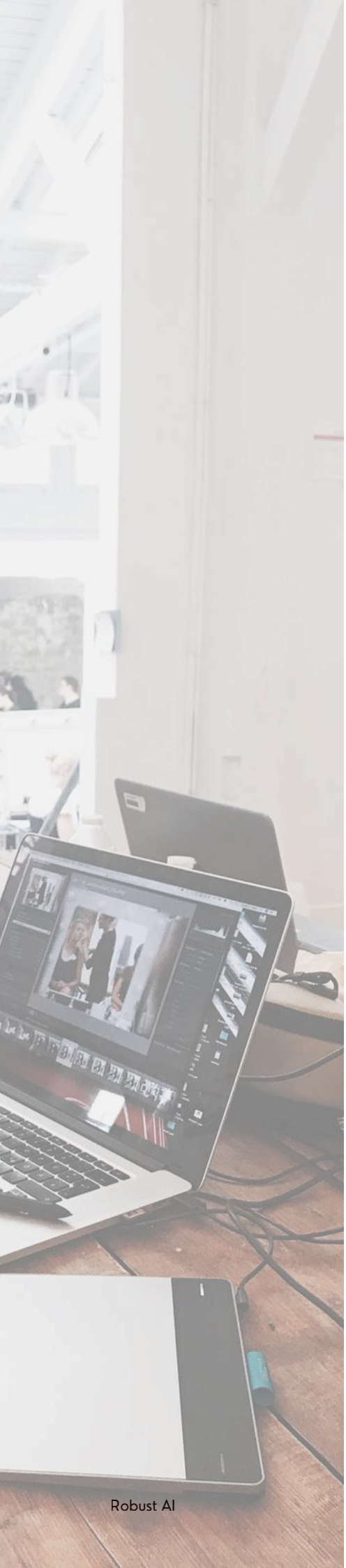
Recent UK incidents illustrate the stakes. The National Cyber Security Centre has warned of adversarial attacks against AI systems in critical infrastructure [17]. Financial services firms have reported attempted manipulation of AI-driven trading systems [18]. Healthcare organisations have discovered vulnerabilities in AI diagnostic tools that could be exploited to cause misdiagnosis [19]. These attacks are not theoretical, but represent active threats that organisations must defend against.

The principle of AI security is straightforward: organisations must protect against attacks that exploit the learning mechanisms of AI systems, not just traditional cyber vulnerabilities. In practice, this requires understanding new attack types, implementing defence-in-depth strategies, and maintaining vigilance as the threat landscape evolves.

What Works: Evidence-Based Security

Leading organisations have adopted several practices to improve AI security, including:

- **Adversarial testing:** Before deployment, AI systems undergo rigorous testing with adversarial inputs designed to cause failure. Red teams specifically trained in AI attack techniques probe systems for vulnerabilities. Organisations establish security thresholds and don't deploy systems that fail to meet them [20].
- **Input validation and sanitisation:** Systems implement multiple layers of input validation to detect and reject potentially adversarial inputs before they reach the AI model. This includes statistical anomaly detection, input bounds checking, and pattern matching against known attack signatures [21].
- **Model hardening:** Techniques such as adversarial training (training models on adversarial examples), defensive distillation, and ensemble methods make models more resistant to attack. While no defence is perfect, hardened models are significantly more difficult to manipulate [22].
- **Supply chain security:** Organisations verify the provenance of pre-trained models, datasets, and tools. They scan for known vulnerabilities, validate model integrity, and maintain inventories of AI components with their security status [23].



UK Evidence and Policy Landscape

The National Cyber Security Centre has published guidance on securing AI systems, emphasising the need for AI-specific threat modelling and defence strategies [24]. The NCSC's AI Security Framework provides a structured approach to identifying and mitigating AI-specific risks.

The UK government's Secure by Design principles are being extended to cover AI systems, with forthcoming guidance expected to mandate security assessments for high-risk AI deployments [25]. Financial regulators including the FCA and PRA have issued guidance on AI model risk management that includes security considerations [26].

ISO/IEC 42001, the international standard for AI management systems, includes security requirements that organisations pursuing certification must meet [27]. Early adopters report that certification drives significant improvements in security practices.

Areas for Investigation and Awareness

Senior leaders and decision-makers should be aware of and investigate:

- **Threat modelling practices:** How comprehensively are AI-specific threats being identified and assessed? Are adversarial attack approaches included in security assessments alongside traditional cyber risks?
- **Adversarial testing capabilities:** What capabilities exist for red-teaming AI systems? Are testing methodologies keeping pace with evolving attack techniques?
- **Supply chain visibility:** What visibility exists into the provenance and security status of AI components? How are third-party models and datasets being validated?
- **Incident response preparation:** Are incident response plans updated to address AI-specific attack scenarios? Do response teams have the expertise to investigate AI security incidents?
- **Regulatory alignment:** How aligned are current practices with NCSC guidance and emerging regulatory requirements for AI security?

RELIABILITY AND PERFORMANCE: ENSURING CONSISTENT OPERATION

The Challenge

AI system reliability differs fundamentally from traditional software reliability. A database query either returns the correct result or fails with an error. An AI system might return a plausible-looking result that is subtly wrong, might perform differently on seemingly similar inputs, or might degrade gradually in ways that are difficult to detect until significant harm has occurred [28].

The evidence on AI reliability in the UK is concerning. Studies have found significant performance variation in AI diagnostic tools deployed across different NHS trusts, with accuracy dropping substantially for patient populations underrepresented in training data [29]. Financial services AI systems have shown performance degradation during market volatility, precisely when reliable predictions are most valuable [30]. Automated decision systems in public services have produced inconsistent outcomes for similar cases, creating fairness and legal compliance challenges [31].

Why does reliability fail? Several factors contribute: training data that doesn't represent the full range of production conditions; testing that focuses on average performance without examining edge cases and failure modes; deployment environments that differ from testing environments; and a lack of ongoing monitoring to detect performance degradation over time.

What Works: Practical Interventions

Organisations achieving reliable AI deployment follow several practices, including:

Comprehensive testing strategies: Beyond measuring average accuracy, robust testing examines performance across subpopulations, stress conditions, and edge cases. Organisations define minimum performance thresholds for all relevant scenarios and don't deploy systems that fail to meet them [32].

Uncertainty quantification: Rather than producing single-point predictions, AI systems are designed to express confidence levels. When the system is uncertain, it can flag outputs for human review or decline to make predictions. This prevents overconfident wrong answers [33].

Continuous monitoring: Production systems are instrumented to continuously measure performance metrics, detect drift, and alert operators to degradation. Dashboards provide real-time visibility into system health [34].

Graceful degradation design: Systems are built to fail safely when AI components underperform. Fallback mechanisms (such as rule-based systems, human escalation, or conservative defaults) ensure that AI failures don't cause catastrophic outcomes [35].

Performance SLAs and governance: Clear service level agreements define acceptable performance, and governance processes ensure accountability for meeting them. When SLAs are breached, defined escalation procedures trigger remediation [36].

UK Evidence and Policy Landscape

The Medicines and Healthcare products Regulatory Agency (MHRA) has established requirements for AI medical device reliability, including clinical validation requirements and post-market surveillance obligations [37]. These provide a model for reliability requirements in other high-stakes domains.

The Financial Conduct Authority's guidance on AI model risk management emphasises the need for ongoing performance monitoring and validation, recognising that initial testing alone cannot ensure reliability [38]. The Bank of England has highlighted AI reliability as a systemic risk concern.

The Alan Turing Institute's research on AI robustness provides technical frameworks for testing and monitoring that organisations can adopt [39]. Industry initiatives are developing benchmarks and standards for AI reliability measurement.

Areas for Investigation and Awareness

Senior leaders and decision-makers should consider:

- **Testing comprehensiveness:** How thoroughly are AI systems tested across different conditions, populations, and edge cases? Are performance thresholds defined and enforced for all relevant scenarios?
- **Uncertainty handling:** Do AI systems express confidence levels? What happens when the system is uncertain? Are there ways to prevent overconfident wrong outputs?
- **Monitoring and alerting:** What monitoring exists for production AI systems? How quickly would performance degradation be detected? What alerting thresholds exist?
- **Graceful degradation:** What fallback mechanisms exist when AI components fail or underperform? How are these fallbacks tested and maintained?
- **Performance governance:** What service level agreements govern AI performance? What accountability mechanisms exist when performance falls below set levels?

DATA INTEGRITY AND QUALITY: BUILDING ON SOLID FOUNDATIONS

The Challenge

AI systems are only as good as the data they're trained on and operate with. Data quality issues that might be minor annoyances in traditional analytics become critical vulnerabilities in AI systems. Corrupted training data can permanently embed incorrect behaviour. Missing data can create blind spots in system capabilities. Stale data can cause systems to operate on outdated assumptions [40].

The evidence on data quality challenges in UK AI deployments is substantial. Healthcare AI systems have shown degraded performance when deployed on data from different electronic health record systems than they were trained on [41]. Financial services AI has been compromised by data quality issues in transaction feeds, causing significant false positive rates in fraud detection [42]. Public sector automated decision systems have produced incorrect outcomes due to inconsistencies in data from different government databases [43].

Data integrity threats extend beyond quality issues to deliberate manipulation. Data poisoning attacks (where adversaries deliberately inject corrupted data into training sets) can cause AI systems to learn incorrect behaviours that are extremely difficult to detect [44]. As organisations increasingly rely on external data sources and crowdsourced labelling, supply chain risks multiply.

What Works: Frameworks for Responsibility

- **Data quality pipelines:** Automated pipelines validate data at every stage from collection through training to inference. Quality checks include completeness validation, consistency verification, freshness monitoring, and anomaly detection [45].
- **Provenance tracking:** Comprehensive lineage tracking documents where data comes from, how it was processed, and who has access. This enables rapid identification of issues and supports audit requirements [46].
- **Data versioning and reproducibility:** Training datasets are versioned alongside models, enabling reproduction of training runs and investigation of issues. Organisations can roll back to previous data versions if problems are discovered [47].
- **Integrity monitoring:** Continuous monitoring detects data drift (changes in data distributions over time) and data quality degradation. Alerts trigger when data characteristics move outside acceptable bounds [48].
- **Supply chain security:** External data sources are vetted for reliability and security. Contracts with data providers include quality guarantees and liability provisions. Critical data dependencies are identified, and alternatives are maintained [49].

UK Evidence and Policy Landscape

The Information Commissioner's Office has issued guidance on data quality requirements for AI systems, emphasising that GDPR's data quality principles apply to AI training data and must be demonstrable [50]. Organisations using personal data for AI must be able to show that the data is accurate, relevant, and up to date.

The UK government's Data Standards Authority is developing guidance on data quality for AI applications in the public sector [51]. This includes requirements for data documentation, quality metrics, and governance processes.

Industry initiatives, including the Open Data Institute's Data Ethics Canvas and the Partnership on AI's data standards work provide frameworks organisations can adopt to improve data governance [52].

Areas for Investigation and Awareness

Senior leaders and decision-makers should consider:

- **Data quality processes:** What systematic processes exist for validating data quality at each stage of the AI pipeline? How comprehensive are quality checks, and what happens when data fails validation?
- **Provenance and lineage:** Can organisations trace data from source through processing to model training? Is this documentation sufficient for audit and incident investigation?
- **Drift monitoring:** What monitoring exists to detect changes in data distributions over time? How quickly would significant drift be detected and addressed?
- **Supply chain risk:** How dependent is the organisation on external data sources? What due diligence has been conducted on data providers? What contingencies exist if data sources become unavailable or compromised?
- **Regulatory compliance:** How aligned are data practices with ICO guidance and emerging regulatory requirements for AI data governance?

OPERATIONAL RESILIENCE: MAINTAINING CONTINUITY

The Challenge

AI systems create new categories of operational risk that traditional business continuity frameworks don't adequately address. When AI is embedded in critical processes, its failure affects not just a single application but potentially cascades across dependent systems and business functions. Recovery from AI failures is complicated by the difficulty of understanding what went wrong and ensuring it won't happen again[53].

UK organisations have experienced significant AI operational failures. Cloud provider outages have disrupted AI services across multiple sectors simultaneously, revealing hidden dependencies [54]. Model updates intended to improve performance have instead caused production failures, with rollback complicated by data state changes [55]. Capacity constraints during peak demand cause AI systems to fail when they are most needed [56].

The concentration of AI capabilities among a small number of providers creates systemic resilience concerns. When organisations depend on the same foundation models, cloud platforms, and tooling, a single point of failure can affect entire sectors. Regulatory bodies have begun treating this concentration as a systemic risk requiring specific attention [57].

What Works: Evidence on Successful Transitions

Comprehensive dependency mapping: Organisations document all dependencies of their AI systems, including infrastructure, data sources, model components, and third-party services. This mapping enables the identification of single points of failure and risks [58].

Redundancy and diversity: Critical AI systems incorporate redundancy at multiple levels, including infrastructure, models, and data sources. Where possible, diversity (using different approaches or providers) reduces the risk of correlated failures [59].

Chaos engineering for AI: Organisations proactively test resilience by deliberately taking down components, corrupting inputs, simulating adversarial attacks, and observing system behaviour. This reveals weaknesses before they cause production incidents [60].

Runbook development and testing: Detailed operational runbooks document procedures for common failure scenarios, including AI-specific incidents like model degradation, adversarial attacks, and data pipeline failures. Regular drills ensure teams can execute these procedures under pressure [61].

Vendor and concentration risk management: Organisations assess and manage their exposure to key AI providers. Exit strategies, alternative suppliers, and contingency plans reduce dependency on any single vendor [62].

UK Evidence and Policy Landscape

The Bank of England and Financial Conduct Authority have designated certain third-party providers as "critical" under the operational resilience framework, with AI cloud providers likely to be included as the framework evolves [63]. Financial institutions must demonstrate they can maintain important business services even if critical AI providers fail.

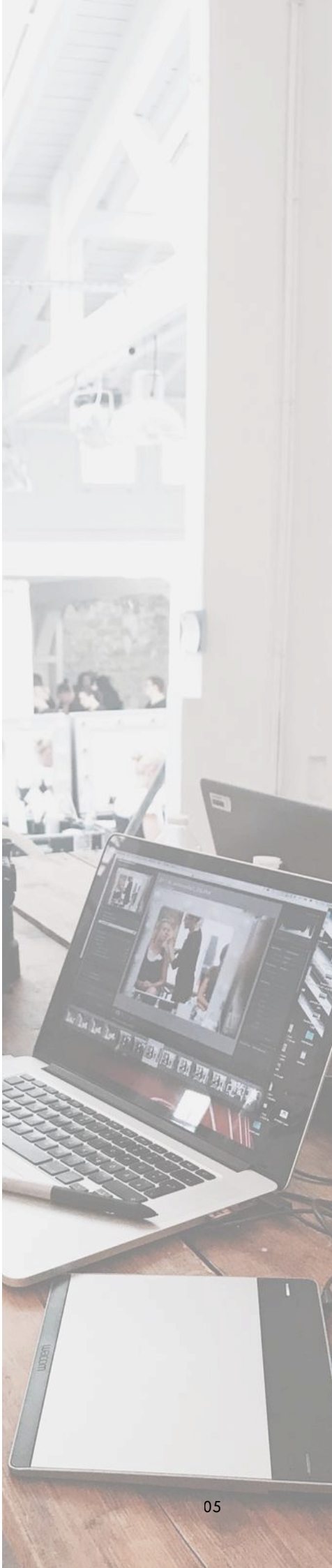
The Digital Operational Resilience Act (DORA), while EU legislation, affects UK financial services firms operating in Europe and provides a model for AI operational resilience requirements [64]. The FCA has indicated it is considering similar requirements for UK-only firms.

NHS Digital's guidance on AI deployment includes operational resilience requirements, recognising that healthcare AI must maintain availability and performance even under adverse conditions [65].

Areas for Investigation and Awareness

Senior leaders and decision-makers should consider:

- **Dependency visibility:** How comprehensively are AI system dependencies documented? Are single points of failure and concentration risks identified and addressed?
- **Resilience testing:** What testing occurs to validate AI system resilience? Are chaos engineering or similar approaches used to proactively identify weaknesses?
- **Incident response capability:** Are operational runbooks updated for AI-specific failure scenarios? Do response teams have the skills and tools to diagnose and remediate AI incidents?
- **Recovery capabilities:** What is the recovery time objective for AI systems? Can organisations actually meet these objectives, and has this been tested?
- **Vendor concentration:** What exposure exists to key AI providers? What contingency plans exist if critical providers become unavailable?
- **Regulatory alignment:** How prepared are organisations for emerging operational resilience requirements that include AI systems?



CONCLUSION: A PATH FORWARD

Robust AI is not a luxury for organisations deploying AI at scale. It is an operational necessity. But the practices described in this framework should be understood as a floor, not a ceiling. They reduce risk; they do not eliminate it. Senior leaders who treat good practice as a guarantee will be caught off-guard. Those who treat it as a starting point and ask hard questions about key risks will be well placed to make responsible decisions.

The four dimensions examined here (security and adversarial resilience, reliability and performance, data integrity and quality, and operational resilience) are interconnected. Security vulnerabilities can cause reliability failures. Data quality issues can create security vulnerabilities. Operational fragility can mask all other problems until a crisis reveals them. Organisations must address all four dimensions systematically.

Two commitments follow from this. First, before any significant AI deployment, organisations should assess the residual risk after mitigations are in place and be willing to conclude that some deployments should not go ahead. Second, monitoring in operation is not optional housekeeping. It is essential because AI systems can develop problems in live use that no amount of prior testing would have revealed.

The evidence is clear: organisations investing in robustness experience fewer incidents, recover faster when incidents occur, and build the trust necessary for AI adoption at scale [66,67]. This investment pays dividends in reduced incident costs, maintained customer confidence, and regulatory compliance.

The UK has an opportunity to lead in robust AI deployment. Our regulatory frameworks are evolving to require robustness. Our research institutions are developing the technical foundations. Our organisations can adopt best practices now, positioning themselves for competitive advantage as AI becomes ever more central to economic activity.

FURTHER READING & RESOURCES

National Cyber Security Centre: https://www.ncsc.gov.uk

Alan Turing Institute: https://www.turing.ac.uk

ISO/IEC 42001: https://www.iso.org/standard/81230.html

Financial Conduct Authority: https://www.fca.org.uk

MHRA: https://www.gov.uk/mhra

UK AI Security Institute: https://www.aisi.gov.uk

REFERENCES -1

1. NHS England. (2025). Artificial Intelligence in Healthcare: Safety and Governance Framework. https://www.england.nhs.uk/digitaltechnology/ai
2. Bank of England. (2024). Financial Stability Report: AI Risks in Financial Services. https://www.bankofengland.co.uk/financial-stability-report
3. Department for Transport. (2025). Connected and Automated Vehicles: Safety Assurance. https://www.gov.uk/government/collections/connected-and-automated-vehicles
4. UK Finance. (2024). AI in Banking: Operational Challenges and Lessons Learned. https://www.ukfinance.org.uk
5. British Medical Journal. (2025). Performance Variation in AI Diagnostic Tools Across NHS Trusts. https://www.bmj.com
6. Public Law Project. (2024). Automated Decision-Making in Public Services: Case Studies. https://publiclawproject.org.uk
7. Marcus, G. & Davis, E. (2024). Rebooting AI: Building Artificial Intelligence We Can Trust. Vintage Books.
8. Goodfellow, I. et al. (2018). Attacking Machine Learning with Adversarial Examples. OpenAI Blog.
9. Geirhos, R. et al. (2020). Shortcut Learning in Deep Neural Networks. Nature Machine Intelligence.
10. Bender, E. et al. (2021). On the Dangers of Stochastic Parrots. Proceedings of FAccT 2021.
11. Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models. Nature Machine Intelligence.
12. Sculley, D. et al. (2015). Hidden Technical Debt in Machine Learning Systems. NeurIPS 2015.
13. Lu, J. et al. (2018). Learning under Concept Drift: A Review. IEEE Transactions on Knowledge and Data Engineering.
14. Wei, J. et al. (2022). Emergent Abilities of Large Language Models. arXiv:2206.07682.
15. Crankshaw, D. et al. (2017). Clipper: A Low-Latency Online Prediction Serving System. NSDI 2017.
16. National Cyber Security Centre. (2024). Threat Assessment: AI-Enabled Cyber Operations. https://www.ncsc.gov.uk
17. NCSC. (2025). Principles for the Security of Machine Learning. https://www.ncsc.gov.uk/collection/machine-learning
18. Financial Conduct Authority. (2024). Market Integrity and AI Trading Systems. https://www.fca.org.uk
19. MHRA. (2025). Safety Alerts: AI Medical Device Vulnerabilities. https://www.gov.uk/drug-device-alerts
20. Microsoft. (2024). AI Red Teaming Best Practices. https://www.microsoft.com/security/blog
21. Papernot, N. et al. (2016). Towards the Science of Security and Privacy in Machine Learning. arXiv:1611.03814.
22. Madry, A. et al. (2018). Towards Deep Learning Models Resistant to Adversarial Attacks. ICLR 2018.
23. MITRE. (2024). ATLAS: Adversarial Threat Landscape for AI Systems. https://atlas.mitre.org
24. NCSC. (2025). AI Security Guidance for Organisations. https://www.ncsc.gov.uk
25. Department for Science, Innovation and Technology. (2025). Secure by Design: AI Systems. https://www.gov.uk/dsit
26. PRA. (2024). Supervisory Statement: Model Risk Management for AI. https://www.bankofengland.co.uk/prudential-regulation
27. ISO. (2023). ISO/IEC 42001:2023 Artificial Intelligence Management System. https://www.iso.org/standard/42001
28. Paleyes, A. et al. (2022). Challenges in Deploying Machine Learning. ACM Computing Surveys.
29. Health Services Research. (2024). AI Diagnostic Tool Performance Variation in NHS Settings.
30. Bank of England. (2024). AI in Financial Services: Operational Risk Considerations.
31. Ada Lovelace Institute. (2024). Automated Decision Systems in UK Public Services. https://www.adalovelaceinstitute.org
32. Breck, E. et al. (2017). ML Test Score: A Rubric for ML Production Readiness. IEEE BigData 2017.
33. Gal, Y. & Ghahramani, Z. (2016). Dropout as a Bayesian Approximation. ICML 2016.
34. Uber Engineering. (2017). Meet Michelangelo: Uber's Machine Learning Platform. https://eng.uber.com
35. Nushi, B. et al. (2017). On Human Intellect and Machine Failures. CHI 2017.

REFERENCES -2

36. Sato, D. et al. (2019). Continuous Delivery for Machine Learning. ThoughtWorks.
37. MHRA. (2024). Guidance on AI as a Medical Device. https://www.gov.uk/mhra
38. FCA. (2024). Machine Learning in UK Financial Services. https://www.fca.org.uk
39. Alan Turing Institute. (2024). Robustness in Machine Learning. https://www.turing.ac.uk
40. Polyzotis, N. et al. (2018). Data Lifecycle Challenges in Production Machine Learning. SIGMOD Record.
41. Ghassemi, M. et al. (2020). A Review of Challenges in Machine Learning for Healthcare. AMIA.
42. UK Finance. (2024). Data Quality in Financial Services AI Systems.
43. National Audit Office. (2024). Data Quality in Government AI Applications. https://www.nao.org.uk
44. Biggio, B. & Roli, F. (2018). Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning. Pattern Recognition.
45. Schelter, S. et al. (2018). Automating Large-Scale Data Quality Verification. VLDB 2018.
46. Herschel, M. et al. (2017). A Survey on Provenance. VLDB Journal.
47. Vartak, M. et al. (2016). ModelDB: A System for Machine Learning Model Management. HILDA 2016.
48. Barash, G. et al. (2019). Monitoring Machine Learning Models in Production. InfoQ.
49. Li, Y. et al. (2021). On the Risks of Machine Learning Supply Chains. IEEE S&P 2021.
50. ICO. (2024). Guidance on AI and Data Protection: Data Quality Requirements. https://ico.org.uk
51. Data Standards Authority. (2025). Data Quality Standards for AI. https://www.gov.uk/dsa
52. Open Data Institute. (2024). Data Ethics Canvas. https://theodi.org
53. Sculley, D. et al. (2015). Hidden Technical Debt in Machine Learning Systems. NeurIPS 2015.
54. Wired UK. (2024). When Cloud AI Goes Down: Lessons from Major Outages. https://www.wired.co.uk
55. Sato, D. et al. (2019). Continuous Delivery for Machine Learning. ThoughtWorks.
56. Google Cloud. (2024). Site Reliability Engineering for ML Systems. https://sre.google
57. Bank of England. (2024). Third Party Concentration Risk in Financial Services AI.
58. Kleppmann, M. (2017). Designing Data-Intensive Applications. O'Reilly Media.
59. Patterson, D. et al. (2002). Recovery Oriented Computing. IEEE Computer.
60. Netflix Technology Blog. (2020). Chaos Engineering Upgraded. https://netflixtechblog.com
61. Google. (2016). Site Reliability Engineering. O'Reilly Media.
62. Gartner. (2024). Managing AI Vendor Concentration Risk. https://www.gartner.com
63. FCA. (2024). Critical Third Parties: Oversight Framework. https://www.fca.org.uk
64. European Commission. (2022). Digital Operational Resilience Act (DORA). https://finance.ec.europa.eu
65. NHS Digital. (2024). AI Deployment Guidance: Operational Resilience. https://digital.nhs.uk
66. McKinsey. (2024). The State of AI: Operational Excellence in AI Deployment. https://www.mckinsey.com
67. Deloitte. (2024). Trustworthy AI: Building Resilient Systems. https://www2.deloitte.com

CONTACT

For more information, or to offer feedback on this report, please contact:

www.DigitalPolicyAlliance.org

info@DigitalPolicyAlliance.org