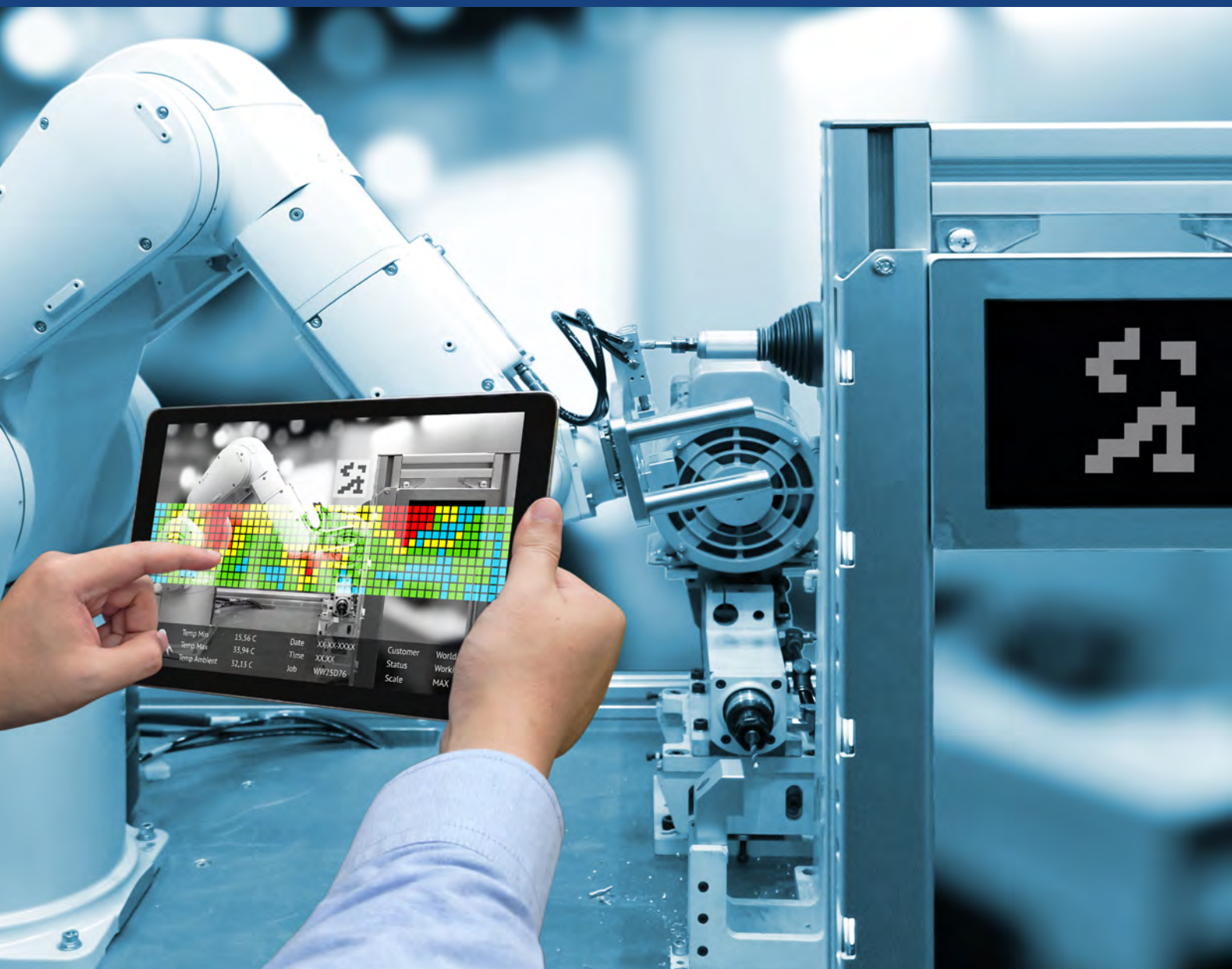
CATAPULT
High Value Manufacturing

Spring 2026

Trustworthy AI adoption

Enabling safe and successful AI adoption in UK industry



Foreword

Artificial Intelligence (AI) has the potential to transform UK manufacturing, driving productivity, resilience and global competitiveness.

However, realising this potential depends not on technological capability alone, but on trust – that AI systems will operate safely, reliably and as intended in complex industrial environments.

Across industry, a consistent message is emerging: uncertainty around AI reliability, safety and governance remains a primary barrier to adoption at scale. Addressing this challenge requires more than guidance or principles; it requires credible, independent mechanisms to test, validate and assure AI systems in real-world conditions.

The case studies in this report provide practical evidence of why third-party AI assurance matters. Drawing on the capabilities of the High Value Manufacturing Catapult, co-led with the National Physical Laboratory (NPL) and delivered through a network of leading organisations, they show how structured, independent evaluation can identify risks early, strengthen confidence and enable safer, faster deployment of AI in manufacturing.

As the Department for Business and Trade's AI Champion for Advanced Manufacturing, I strongly support the development of a coordinated national approach to AI validation and assurance. Embedding these practices into the industrial AI lifecycle is essential if the UK is to unlock the full value of AI while maintaining public trust and international leadership in responsible innovation.

Professor Chris Dungey

HVM Catapult CTO and AI Champion for Advanced Manufacturing

The High Value Manufacturing (HVM) Catapult helps drive economic growth across the country by keeping UK manufacturing innovative, productive and globally competitive.

Each year, thousands of businesses come to HVM Catapult seeking solutions to a particular problem or looking for ways to improve the products they sell, the way they make them and the skills of their workforce which could help them attract inward investment and prosper in the global marketplace.

HVM Catapult's national network of centres and 3,800 employees provide those companies with access to world-class facilities and expertise that would otherwise be out of reach.

Contents

Foreword	2
Introduction	4
AMRC: Bridging innovation and trust in AI-enabled inspection systems	6
CPI: Validating federated learning models for high-temperature manufacturing	9
MTC: LLM-powered robotics for safety-critical site inspections	13
NMIS: Verification of a bolt detection system for trustworthy AI	16
NCC: Assuring trustworthy AI in NDT: insights into generalisation challenges	20
WMG: Thickness prediction model for battery electrode in manufacture	24
Conclusion and recommendations	27
Acknowledgements	27

Introduction

The application of AI is rapidly reshaping global industry. Yet trustworthiness in AI is essential for its adoption, which hinges on transparency, reliability, ethical, secure and accountable AI development.

Across the UK, industrial organisations consistently report that uncertainty around AI reliability, safety and performance remains one of the primary barriers to deployment at scale.

The UK Government's AI assurance roadmap[1] emphasises that independent verification is fundamental to enabling adoption, encouraging investment and ensuring AI systems 'work as intended'. Government analysis reinforces this point: the UK AI assurance market, valued at £1.01bn in 2024, is projected to grow to £18.8bn by 2035[1] if barriers to responsible AI adoption are removed. This growth will only be realised if industry can rely on credible, independent mechanisms to validate AI at system level.

This document explores key case studies that demonstrate, through real industrial examples, why third-party AI testing and validation are essential for safe and confident AI adoption, and how to archive it. Each case study highlights specific challenges encountered when evaluating AI systems, ranging from functional accuracy to robustness, governance, explainability and real-world performance, and shows how structured, independent assessment can reveal failure modes and risks that are otherwise difficult to detect. They also illustrate the value gained through rigorous AI assurance, avoiding catastrophic losses incurred by the deployment of under-developed and unsafe AI solutions.

The findings show that evaluating AI requires looking beyond performance alone, ensuring all pillars of trustworthy AI are addressed, Figure 2. Building on the MTC's Trustworthy AI whitepaper[3] this work expands those pillars to deliver a more complete validation framework aligned with industry needs, as summarised in Figure 1.

These case studies draw on expertise from across the HVM Catapult, which is leading a national AI validation network[2], bringing together leading organisations to provide coordinated testing infrastructures and methodologies that support trustworthy AI adoption across UK manufacturing. This collaboration with organisations including the Alan Turing Institute and National Physical Laboratory reinforces the strategic importance of developing robust assurance practices within UK industry.

Collectively, the case studies contained in this compilation demonstrate in practical, measurable terms why third-party AI assurance is indispensable. They reveal recurring risks, systemic failure modes and significant value-creation opportunities that cannot be uncovered without structured, independent evaluation leading to better insights and recommendations to support successful AI adoption within your business.

Although each study represents different technologies and sectors, together they make a compelling case: robust, transparent AI validation is foundational to accelerating safe, competitive and widespread AI adoption across UK manufacturing and beyond.

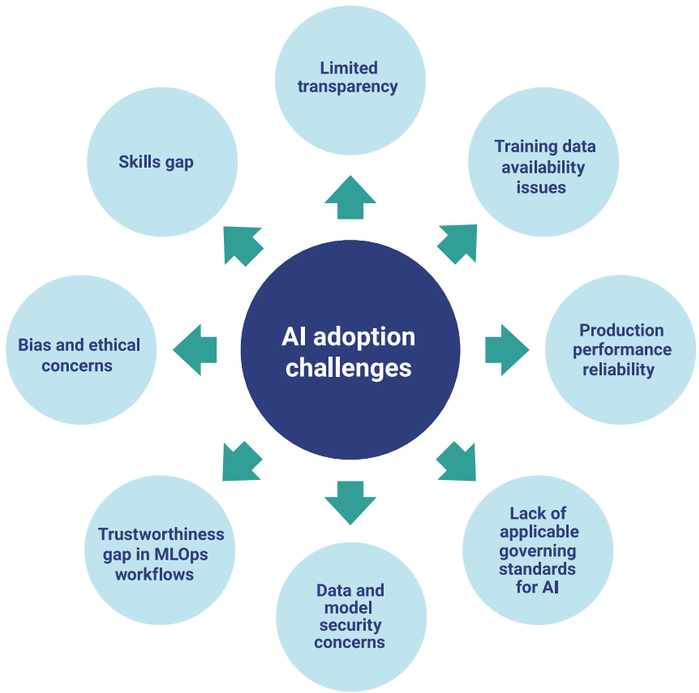


Figure 1 AI adoption challenge in the manufacturing industry.[2]

Figure 2 AI assurance pillars.

[1] https://www.gov.uk/government/publications/trusted-third-party-ai-assurance-roadmap/trusted-third-party-ai-assurance-roadmap
[2] https://hvm.catapult.org.uk/news/unlocking-the-full-potential-of-ai-adoption/
[3] https://mtcprod.s3.eu-west-1.amazonaws.com/s3fs-public/2024-07/Trustworthy%20AI%20Framework.pdf

AMRC: Bridging innovation and trust in AI-enabled inspection systems

Overview

Composite manufacturers currently rely on manual visual inspection or basic automated systems that struggle with the complex, textured nature of woven fabrics. Users need a solution that can detect subtle defects at production speeds to reduce scrap rates and ensure structural integrity. Woven composite materials are critical in industries requiring high strength-to-weight ratios, specifically aerospace, automotive and wind energy.

The system is a computer vision solution designed to automate the inspection of woven composite fabrics. To handle the high variability of fabric textures without requiring scarce labelled defect data, only unsupervised anomaly detection models are under consideration, specifically PatchCore.

- **Model type:** Anomaly detection
- **Data type:** Images.

Why AI assurance testing is needed:

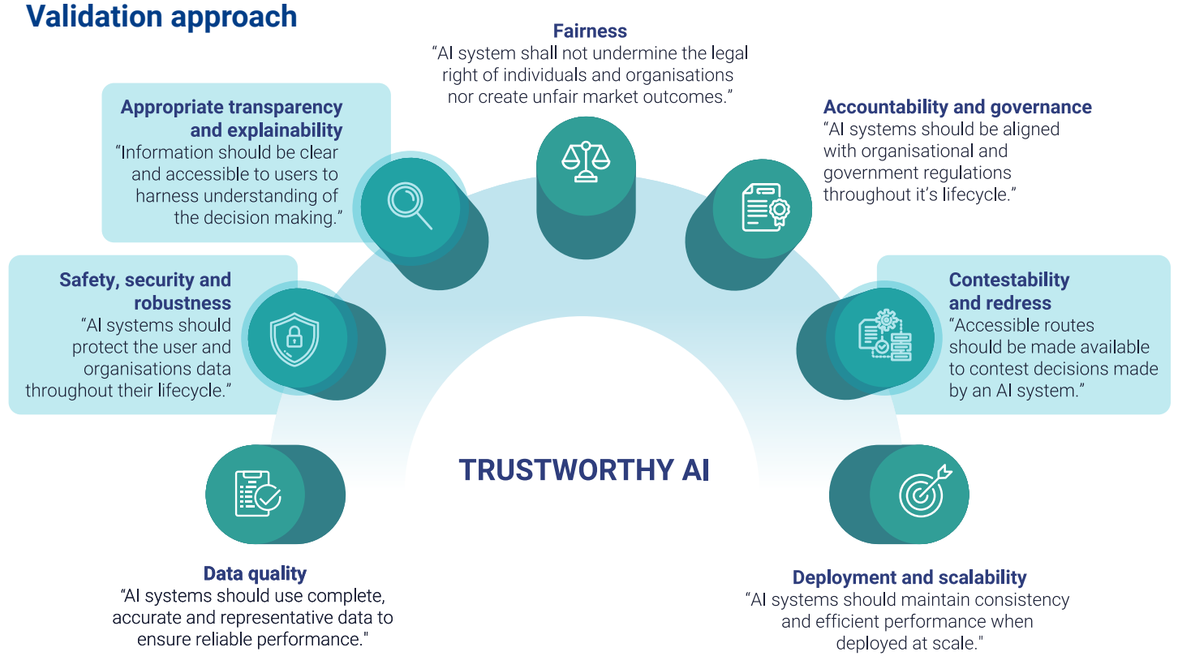
Safety impact: Missed defects can pose significant safety concerns in both the final product and further along the production line.

Financial impact: Identifying defects in composite materials is a vital part of quality control and assurance, impacting company reputation and sales. Incorrect identification of anomalies could also lead to an increase in cost due to scrap rates.

Mitigating unsupervised model risk: Without assurance testing to mathematically define the "normality" threshold, the system risks hallucinating defects or failing to flag novel defect types it hasn't encountered before.

Regulatory compliance: Assurance testing creates the audit trail necessary to demonstrate that the model's decisions are based on valid physical features, e.g. fibre breakage, rather than artifacts or noise, ensuring the system is auditable by external regulators if necessary.

Validation approach



Impact

Component tested	Score	What this translates to
Performance testing: Verify whether the model can accurately and reliably identify images of defective samples. (Image-level AUROC*).	96%	**MONITOR** – While the AUROC* score is encouraging and suggests high levels of model performance, further tests in production environments with much larger scale data would be required before deployment.
Explainability testing: Compare the localisation of the model's predicted heatmaps to the ground truth. This assesses the "reasoning" of the model for its decisions and therefore the quality of its learning (Pixel-level AUROC*).	~81%	**ADDRESS** – This score tends to be significantly lower than the image level score, suggesting that the model may not be localising anomalies accurately. This suggests that the model may not be learning as intended and may be using other unknown factors to influence its decision making.
Flexibility testing: Assess the model's response to new "normal" images that represent a range of acceptable qualities or a change in production.	N/A	**ADDRESS** – The model reacts very poorly to new examples of "normal" without retraining or refinement.

The area under the receiver operating characteristic (AUROC) curve measures a model's ability to distinguish between binary classes (defective vs non-defective).

Benefits of testing

- **Exposure of 'right answer, wrong reason' risks:** Testing prevented the deployment of a model that may have been relying on background noise or image artifacts to flag defects rather than the defect itself. This insight mandated the shift to a human-in-the-loop (HITL) workflow, ensuring operators verify the specific region of interest rather than blindly trusting the model's output
- **Identification of operational fragility:** This prevented a potential production stoppage. Had the system been deployed 'as is', a standard change in material batch could have triggered a mass false-positive event. The testing highlighted the critical need for a continuous integration/continuous deployment (CI/CD) pipeline to rapidly retrain the model on new 'normal' data before it hits the line
- **Data-driven governance and threshold setting:** Instead of guessing a sensitivity threshold, the testing provided empirical data to set the optimal anomaly score cutoff. This balances the trade-off between scrap rate and defect escape rate, allowing the business to tune the model according to their criteria
- **Justification for infrastructure investment:** The results clearly justified the recommendation to integrate MLflow. The testing proved that because the model is sensitive to data drift, rigorous lifecycle management and parameter tracking are not optional "nice-to-haves," but essential requirements for maintaining the detection rate over time.

Recommendations

- **Human-in-the-loop interface:** This system lends itself well to a human-in-the-loop interface where anomaly scores close to the pre-determined optimal threshold (above or below) are flagged. These flagged examples can then be reviewed by a human operator using the heatmap. Further to this, false positives could then be added to the memory bank of the model
- **Training improvements:** MLflow can be integrated into PyTorch so could be used for traceability during training to track learning metrics, model performance, hyperparameters and other relevant parameters, thereby supporting reproducibility and systematic experimentation.

CPI: Validating federated learning models for high-temperature manufacturing

Overview

The study integrated a federated learning (FL) platform into an advanced materials manufacturing challenge. It is characteristic of a problem where data is not shareable due to privacy concerns between commercial companies but where building a global model and sharing that would be beneficial to all. This materials challenge focuses on the use and optimisation of high temperature furnaces, where CPI uses them for battery cathodes and partners for e.g. ceramic parts.

AI system description: The application focuses on using FL to build a global predictive model without sharing sensitive datasets. The aim is to demonstrate that FL enables the construction of higher-quality predictive models than would be achievable using data from using single device in isolation.

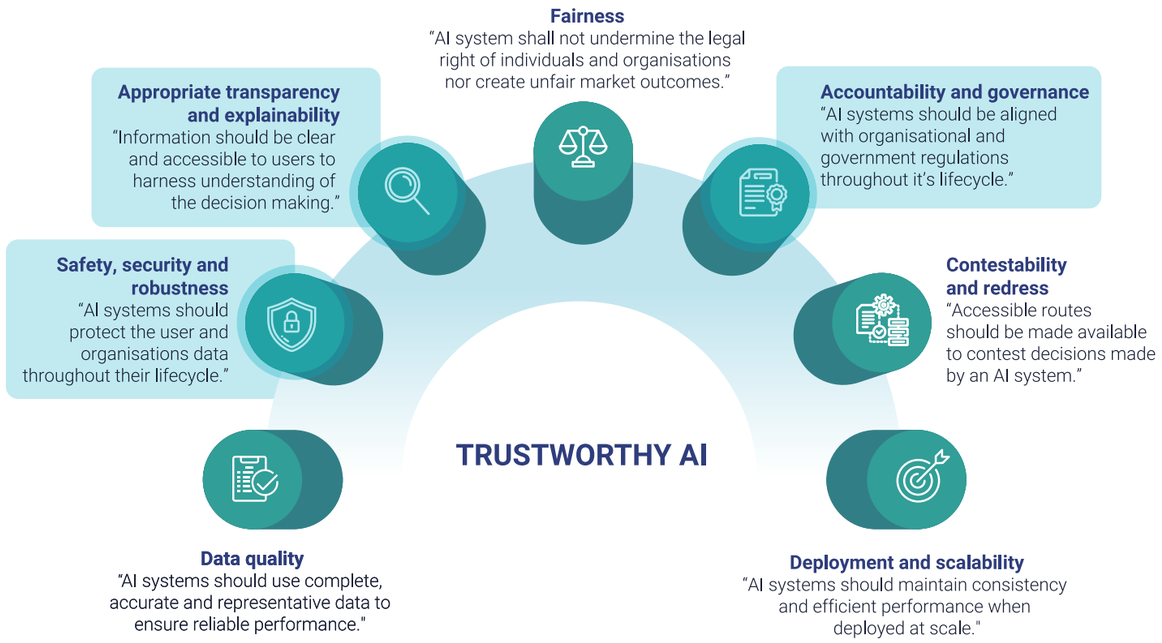
- **Model type:** Supervised regression, XGBoost, deployed in a federated learning framework
- **Data type:** Categorical (sample, furnace); continuous quantitative time series (temperature); continuous quantitative (XRD peak ratio, mass/g, ramp rate/°C min-1); ordinal (quality score/1-5).

Why AI assurance testing is needed:

Our federated system models are intended for distribution across high-impact manufacturing in furnaces. Companies benefit from contributing to a global model, but trustworthiness must be maintained across the federated infrastructure for adoption to work. For example:

- **Trust and adoption:** Raw data is never centrally inspected, and stakeholders have different furnace configurations. Users need confidence that models behave predictably. Without this confidence, users or sites will not accept the FL outputs
- **Model robustness and reliability:** Federated learning systems must produce consistent predictions across sites and data distributions. Validation should ensure resilience to noise, missing data, and domain shifts, and highlight which features drive predictions, supporting safe operational use
- **Safety and financial impact:** Lack of validation affects both the ROI from the FL system and safety when compared to an unfederated equivalent. Errors or hidden biases will lead to reduced product yields and quality, increased energy costs or even risk furnace damage
- **Data security, privacy and governance:** FL relies on datasets from multiple partners spread across sites in a system that promises data will not be shared. The system must demonstrate that data is securely protected from security breaches and that the data quality is maintained.

Validation approach



Impact

Component tested	Score	What this translates to
Model explainability: We investigated model explainability by using Shapley Additive Explanations (SHAP) on two separate but similar datasets.	72%	**MONITOR** – The D-Shap semantic fidelity score between datasets was 0.0141. A good fidelity score is <0.05, above this value it would be a high-risk model and poorly explainable. Despite this score we must put it in context against other aspects of the model to address more fundamental issues. Our outcome is to monitor this value while improving the quantity and quality of datasets collected and continuously reassess.
Accountability and governance: We investigated the platform's current capacity for version monitoring, model tracking and change logging and compared this capability to the model cards toolkit.	66%	**MONITOR** – In our test environment models can easily be tracked using the FL platform's built in system tracking key data types. Comparison across different versions is possible but not easy which will make this difficult for operators and should be addressed if it becomes a significant burden.
Resistance to data attacks: We leveraged the embedded adversarial robustness toolbox within our federated learning platform to evaluate the impact of model targeted attacks through poisoning and Byzantine methods.	50%	**ADDRESS** – The system successfully detected out of range changes in the model caused by inputs from poisoning and Byzantine attacks. When deviations were detected, the system alerted users, paused learning on the affected data and prevented these from compromising the global model. The sensitivity to these attacks needs to be further interrogated to understand the vulnerability of the system and for this to be appropriately set-up.
Malicious device recognition: We investigated the platforms ability to detect when a device changes modelling parameters to benefit itself or damage overall model performance.	50%	**ADDRESS** – The platform automatically stopped training and flagged when a device maliciously changed η, the learning rate. Further testing is needed to appropriately set up this across the federated system as the datasets expand.

Benefits realised by testing

- **Reduced unplanned downtime:** Furnace drift and failed runs result in hours of lost production per event. By detecting drift earlier and harmonising furnace behaviours with federated learning, we expect to reduce the associated downtime significantly
- **Financial impact:** Improved furnace consistency and prediction accuracy is expected to reduce scrap and rework, lowering the number of runs required to meet manufacturing goals. With a 10% improvement scaling to tens of thousands of pounds per furnace per annum at a production scale
- **Reduced operational risk:** By properly configuring adversarial detection and model version control, the system prevents attacks or unfair updates from biasing the model, while providing operators with transparency over operations. With these safeguards in place, losses of several thousand pounds per production line per year can be avoided, making this a critical to realise financial gains from the system.

Recommendations

The system requires improvement before deploying into a live environment.

- **Improve model accuracy:** The RMSE of our models is too high, 35 – 46% showing that while SHAP result highlights the relationship between features, raw predictions are not reliable enough to guide production. Prioritise expanding the datasets and improving model calibration to reduce error to <15 – 20% RMSE before live deployment
- **Enhance process consistency:** Literature highlights the importance of consistency in furnace behaviour, both across different furnaces and within their heating zones. Based on this we target manufacturing conditions that achieve >90% consistency in R&D, and >95% in production, exceeding typical industry expectations for predictable performance
- **Define robustness thresholds for attacks:** Define and proceduralise the correct detection thresholds for poisoning and Byzantine attacks by benchmarking the system sensitivity using controlled scenarios to create meaningful outputs
- **Strengthen model governance and operator usability:** Improve model versioning with clearer comparison, traceability and deployment audit trails. Align governance with model cards to enable cross-business transparency and regulatory readiness, reducing operational friction and improving trust in the system deployment.

MTC: LLM-powered robotics for safety-critical site inspections

Overview

Inspection in sectors like construction, rail and agriculture is critical for ensuring safety, compliance and operational efficiency. These inspections detect structural issues, prevent costly failures and reduce risks to human life. Traditionally, inspections are manual, time-consuming and expensive, often requiring skilled personnel to operate in hazardous environments. To address these challenges, organisations are exploring AI-driven robotic solutions for remote inspection.

A solution was developed using large language models (LLMs) integrated with a mobile robotic platform. Through a simple web interface, users can issue natural language commands to control robots remotely.

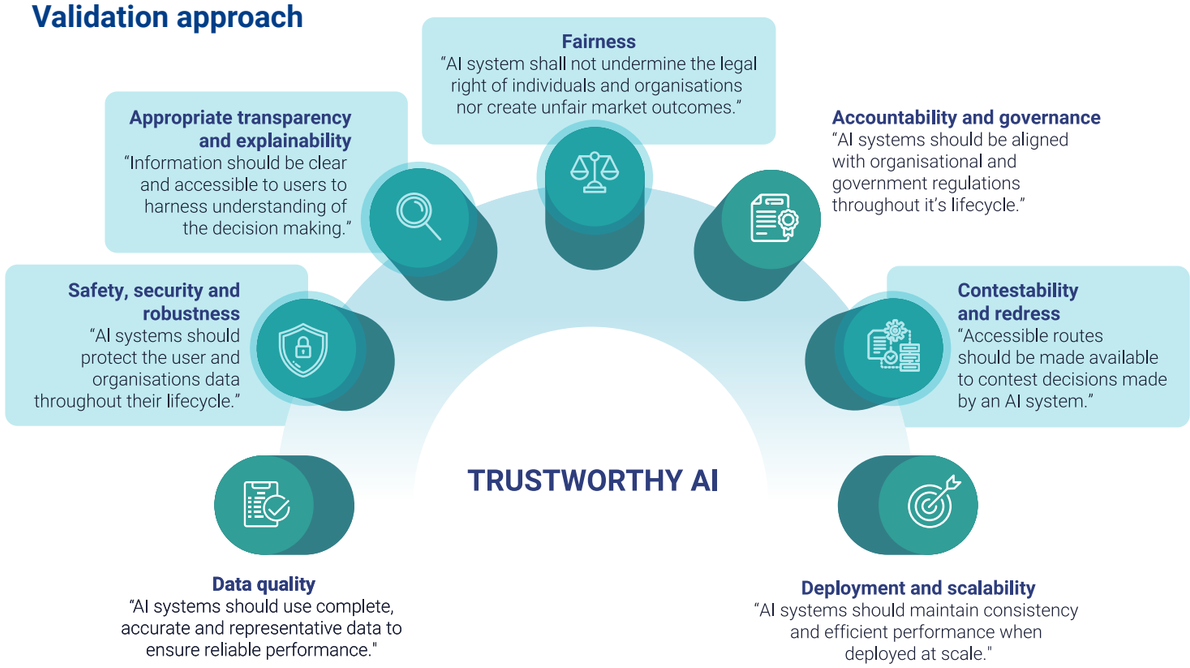
- **Model type:** LLM (Natural language processing)
- **Data type:** Text and images.

Why AI assurance testing is needed:

Adopting this AI system without thorough validation can lead to operational failures, safety hazards and financial losses. For example:

- **Safety impact:** Incorrect robotic commands during inspection could cause accidents, damage assets, or endanger personnel
- **Financial impact:** Unplanned downtime in sectors like rail or construction can cost thousands of pounds per hour. A single failure could lead to penalties of up to £100,000 per incident, reputational damage and increased insurance premiums
- **Regulatory compliance:** Many industries require demonstrable evidence of system reliability before approval for use
- **Trust and adoption:** Stakeholders need confidence that AI systems behave predictably under all conditions, including ambiguous or invalid inputs, before committing to production.

Validation approach



Impact

Component tested	Score	What this translates to
Functional testing: Verify whether the AI consistently produced correct outputs for valid commands.	88%	**MONITOR** – While this suggests the system is generally predictable, it falls short of the reliability expected for production in safety-critical sectors. At this level, unexpected behaviours remain a significant risk, potentially causing downtime costs of £10k - £50k per hour.
Consistency testing: Measure repeatability and stability of responses under varied conditions.	65%	**ADDRESS** – 35% incorrect outputs could lead to operational errors if unchecked. Without rigorous testing, these errors could cause asset damage or safety hazards, potentially millions in liability. The model needs further improvements for general functionality.
Ambiguity testing: Assess robustness to unclear or linguistically diverse prompts, touching on fairness and inclusivity.	17%	**ADDRESS** – Poor handling of unclear commands means operators need clear input guidelines. The system is likely to fail with linguistic diversity, e.g. different dialects or grammar inaccuracies.
Error handling: Ensure the system safely managed invalid or nonsensical inputs without generating hazardous commands.	81%	**MONITOR** – Good ability to reject invalid inputs, but not strong enough for production. However, it still poses meaningful safety risks, as the system may occasionally produce commands when it should not. These gaps need closing before deployment in hazardous or regulated environments.

Benefits realised by testing

- **Risk reduction:** Identified critical failure modes, e.g. poor handling of ambiguous input, before production, avoiding potential safety incidents and costly downtime
- **Financial impact:** Prevented uncontrolled deployment that could lead to asset damage or operational delays, potentially saving £100k+ per incident
- **Informed decision making:** Provided stakeholders with quantified performance metrics, e.g. 88% functional and 65% consistency accuracy, to guide safe adoption or further development
- **Development pathways:** Highlighted areas for improvement, e.g. error handling and ambiguity, to optimise future versions.

Recommendations

- **Human-in-the-Loop safeguards:** Until reliability improves, ensure all commands are supervised or approved by an operator to mitigate operational and safety risks
- **Improve functional accuracy:** Prioritise model refinement to reduce the current 35% error rate in valid commands. This is essential to prevent asset damage and operational disruption
- **Enhance consistency:** Target >95% consistency to meet industry expectations for predictable behaviour and reduce risk of downtime-related financial losses
- **Address ambiguity handling:** Implement robust input guardrails, controlled vocabularies or pre-processing layers to prevent misinterpretation of unclear or linguistically diverse commands
- **Strengthen error handling:** Close remaining gaps so the system reliably rejects invalid inputs >95% of the time, avoiding unsafe or unintended outputs.

NMIS: Verification of a bolt detection system for trustworthy AI

Overview

Assembly validation in manufacturing environments is critical for ensuring product quality, safety and operational efficiency. In low-volume, high-mix production scenarios even minor assembly errors can have catastrophic consequences. Traditional manual inspection methods are prone to human error, inconsistent and often fail to catch defects in real-time. To address these challenges, Thales, in partnership with NMIS through an Innovate UK initiative, has developed an AI-driven computer vision solution for real-time assembly validation.

AI system description: A solution was developed using advanced computer vision technology deployed on edge computing hardware. Through a depth camera system, the AI monitors assembly processes in real-time and validates component placement.

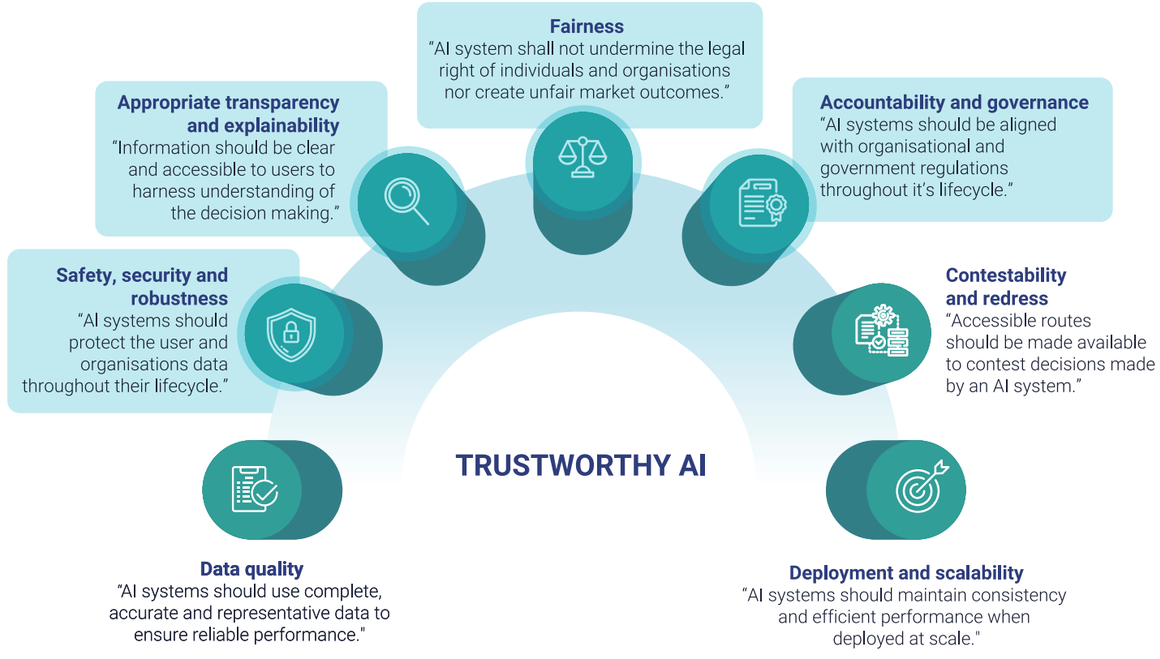
- **Model type:** EfficientDet (scalable and efficient object detection)
- **Data type:** Visual data (RGB-D images from depth camera at up to 100 FPS).

Why AI assurance testing is needed:

Adopting this AI system without thorough validation can lead to operational failures, safety hazards, and financial losses. For example:

- **Safety impact:** Undetected assembly errors could lead to component failures in critical aerospace or defence applications, potentially endangering lives and causing catastrophic equipment failures
- **Financial impact:** Production delays due to false positives or undetected defects can cost manufacturers thousands of pounds per hour. Defective components reaching customers could result in expensive recalls, warranty claims and severe reputational damage
- **Regulatory compliance:** Aerospace and defence manufacturing requires rigorous quality assurance documentation and demonstrable evidence of system reliability to meet industry standards and safety certifications
- **Trust and adoption:** Manufacturing operators need confidence that the AI system performs reliably across diverse operational conditions, including varying lighting, component orientations and potential adversarial scenarios, before integrating it into production workflows.

Validation approach



Impact

Component tested	Score	What this translates to
Explainability testing: Verify whether the AI focuses on relevant visual features (bolts) rather than spurious background correlations using SHAP analysis.	100%	**MAINTAIN** – The model demonstrates excellent transparency, with saliency maps showing SHAP values >0.65 strongly correlating with actual bolt positions.
Accountability and documentation: Assess completeness of model documentation including capabilities, limitations, and operational boundaries using model card toolkit.	90%	**MONITOR** – Comprehensive documentation provided strong transparency into model development and performance benchmarks (81.9% mAP@0.5). However, critical operational boundaries (required lighting ranges, exposure limits, contrast thresholds) are not documented.
Security and robustness testing: Ensure resilience against adversarial attacks and intentional manipulation using adversarial robustness toolbox (ART).	71%	**ADDRESS** – While the model shows excellent robustness to noise (100%) and patch attacks (100% at ε=0.01), it has critical vulnerabilities to optimisation-based gradient attacks: 66.7% accuracy drop under C&W attacks and 33.3 – 66.7% drops under PGD attacks at low perturbation levels.
Fairness testing: Measure consistent performance across diverse operational conditions including lighting, sharpness, contrast and exposure variations using DeepChecks.	81%	**ADDRESS** – Severe performance degradation under real-world conditions: 83% drop in bright lighting, 98% drop with normal sharpness images, 26% drop with overexposure and 28% drop with normal contrast.

Benefits realised by testing

- **Risk reduction:** Identified critical failure modes before production, preventing deployment of an unsafe system that could allow defective aerospace/defence components to reach customers
- **Financial impact:** £200,000 – £300,000 saved through early identification of vulnerabilities
- **Informed decision making:** Provided stakeholders with quantified performance metrics establishing clear verdict: system NOT safe for production without significant improvements
- **Development pathways:** Highlighted priority improvements to enable safe deployment: retrain with diverse background/ environmental conditions, implement adversarial defences (C&W/PGD), complete operational documentation with boundaries and deploy human-in-the-loop safeguards.

Recommendations

- **Robustness against diverse background/ environmental conditions (Fairness):** Retrain model with diverse training data covering full lighting spectrum, varied sharpness levels, and multiple exposure/contrast ranges. Target >95% performance consistency across all operational conditions
- **Implement adversarial defences (Security):** Deploy adversarial training targeting C&W and PGD attack patterns, implement input validation and anomaly detection and establish real-time monitoring for systematic accuracy degradation
- **Complete operational documentation (Accountability):** Document explicit operational boundaries including required lighting ranges, exposure/contrast thresholds, camera positioning requirements, known failure modes and escalation procedures.

NCC: Assuring trustworthy AI in NDT: insights into generalisation challenges

Overview

Quality control in manufacturing is critical for companies that produce large volumes of composite parts and must meet stringent inspection requirements. This is why advanced techniques such as 3D-ultrasound defect detection are integral part for identifying flaws accurately and non-destructively. Detecting defects in A-scan ultrasound data is traditionally a manual process that requires extensive operator expertise and is time-consuming. This motivates the use of machine learning (ML) techniques to automate defect detection and improve efficiency, consistency, and accuracy.

AI system description: A standalone AI tool that reads single-channel A-scan ultrasound measurements and automatically tags potential defects with (a) defect presence, and (b) estimated defect depth/size. The tool provides consistent, auditable annotations that can be exported to standard IIoT for NDT QA workflows. The results were written into MS Word report with constructed C-Scans for location and depth and summary by size of the individual flaws.

The system uses a hybrid machine learning approach combining signal processing, time-series modelling, and deep learning. The core models are a classification model: GRU (Gated Recurrent Unit), which is handling sequential data and capturing temporal dependencies and a depth regression model: 1D Convolutional Neural Network 1D-CNN, for extracting spatial patterns from sequences. Model optimisation is performed using Hyperopt for efficient hyperparameter tuning, and Bayesian Active Learning techniques are applied to prioritise informative samples during training. MLflow is used for platform-independent experiment tracking and model version control.

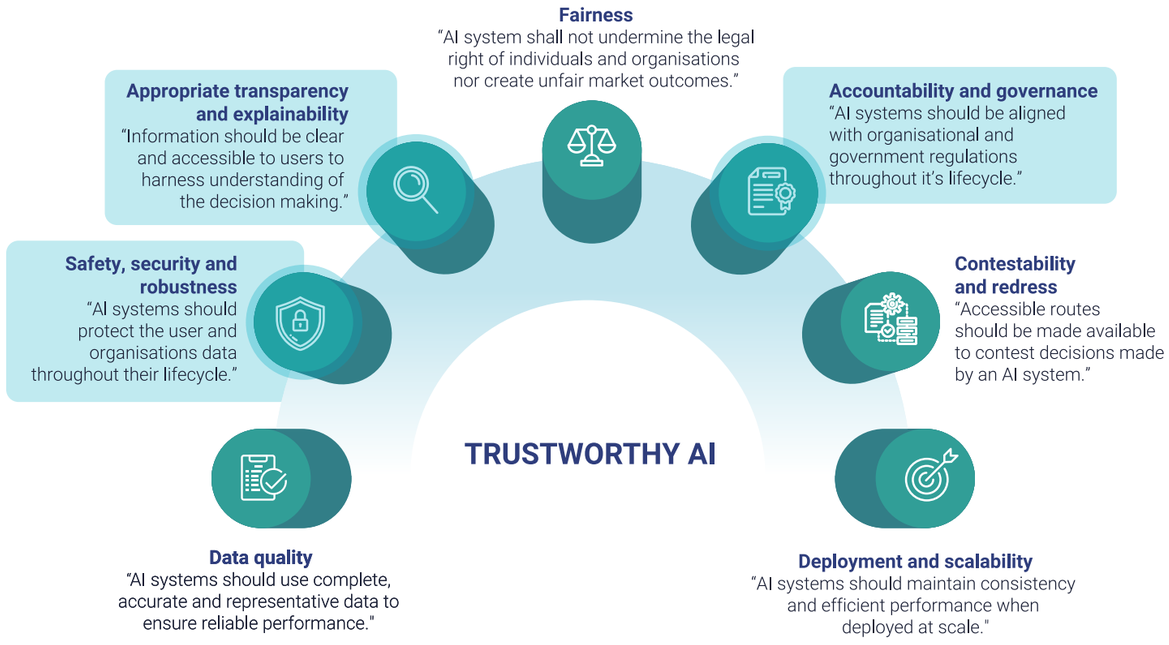
Why AI assurance testing is needed:

AI assurance testing is essential in NDT because the technology directly supports safety-critical decisions – decisions that can affect structural integrity, operational safety and regulatory compliance. Ensuring that AI behaves reliably, transparently, and consistently is not optional, it is fundamental.

- AI systems in NDT help detect cracks, corrosion, delamination, weld defects, and other anomalies
- Any AI failure – missed defects or false alarms – can lead to safety hazards, equipment failure, costly shutdowns, regulatory breaches
- NDT environments vary widely due to materials, geometries, sensors and noise levels. Without assurance testing, models may: overfit to specific datasets, misinterpret new inspection conditions, produce overconfident yet incorrect predictions
- NDT inspectors must understand why the AI is flagging a defect. Assurance testing verifies that the model bases its decisions on physically meaningful features, attribution aligns with domain knowledge, and operators can interpret and contest AI outputs
- Sectors such as aerospace, nuclear, energy and infrastructure require stringent validation of any system influencing inspection results.

AI Assurance testing provides documented evidence that the model meets these compliance expectations.

Validation approach



Impact

Component	Key findings	Implication
Transparency and explainability	Documentation and explainability tools work effectively; model reasoning aligns with signal characteristics.	**MAINTAIN** – Framework is sound, but interpretability does not compensate for limited generalisation.
Robustness and generalisation	Despite multiple dataset splits and attribution checks, the model fails to generalise beyond the specific composite ply used during training.	**ADDRESS** – Broader dataset collection and re-training required; current model not suitable for other part types.
Feature attribution	Positive contributions correspond to defect peaks; negative to normal material signatures.	**MAINTAIN** – Confirms model behaviour is domain-consistent.
Safety and confidence thresholding	Thresholds and human-in-the-loop controls operate correctly.	**MONITOR** – These safeguards remain essential given weak generalisation.
Governance and traceability	MLflow versioning and clear accountability are in place.	**MAINTAIN** – Supports controlled iteration while addressing generalisation gaps.
Human-centric design	Model supports operators but cannot be relied on for other materials.	**ADDRESS** – Communicate limitations clearly to users to avoid over-trust.

Benefits realised by testing

- The framework has not 'proved the model is ready' but it has proved the process is safe – by detecting poor generalisation early and enforcing the controls needed to iterate responsibly. It made weaknesses visible, measurable, and actionable
- Used at design-time, it ensures the model, metadata, and explainability outputs are structured to support governance and reproducibility.

Recommendations

A robust and trustworthy testing framework is essential not only for evaluating models after development but also for shaping how models are built from the very beginning. This framework should serve two equally important functions:

1. **Foundation for trustworthy model development**
   It provides the principles, requirements and structure needed to ensure that models are designed with trustworthiness in mind from the start – covering transparency, safety, explainability and governance as core design objectives
2. **Ongoing verification within the pipeline**
   The same framework should also act as an integrated checking mechanism within the MLOps pipeline. This means models must be prepared and structured in a way that allows them to be automatically tested, validated and monitored against these trust criteria throughout their lifecycle

- By using the framework both as a design guide and as a continuous verification system, organisations ensure that trustworthiness is not an afterthought, but a built-in property of every model iteration – from conception to deployment.

WMG: Thickness prediction model for battery electrode in manufacture

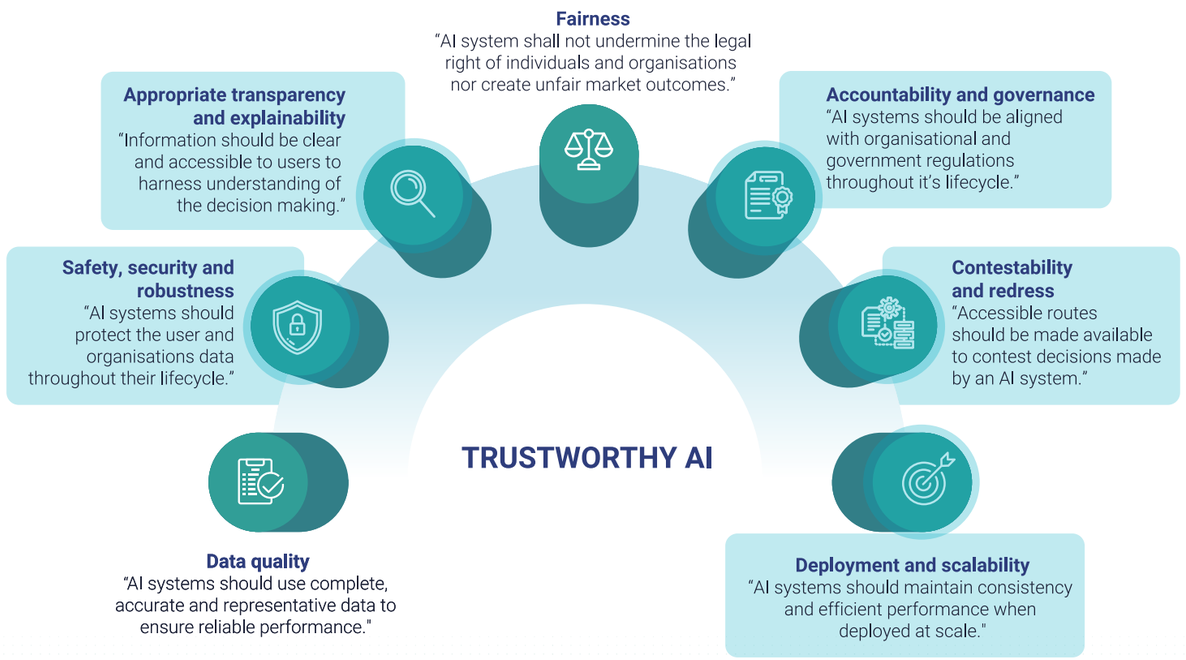
Overview

Battery manufacturing involves several stages, with calendaring being a key step for ensuring electrode thickness uniformity. Existing in-line monitoring methods, such as mechanical calipers and optical sensors, are limited by surface sensitivity, lack of continuous measurement and high implementation costs. This project proposes a novel ultrasound-based sensing approach combined with a tailored deep neural network to address these limitations.

AI system description: A deep neural network was designed using swin-transformer and inception-1D-based architectures for electrode thickness prediction using ultrasonic signals and process parameters.

- **Model type:** Regression
- **Data type:** Ultrasound signals and tabular data (process parameters).

Validation approach



Impact

Component tested	Score	What this translates to
Functional testing: Verify correct input handling, deterministic behaviour under controlled conditions, and numerically stable, physically plausible predictions.	100%	**MAINTAIN** – All sub-tests passed, including schema and data-integrity validation, deterministic behaviour under controlled conditions, stable predictions under boundary conditions and strict monotonicity with respect to nip-pressure. Results were numerically stable, physically plausible and fully traceable.
Robustness testing: Assess resilience to realistic degradations such as sensor noise, missing UT data, timestamp drift and distribution shift.	50%	**MONITOR** – The model showed strong resilience to sensor noise and acceptable tolerance to temporal lag, but failed severely under missing UT data, and OoD detection was unstable due to high variability in Mahalanobis scores. Robustness is adequate for prototyping but insufficient for deployment without mitigation strategies.
Latency and performance testing: Confirm suitability for real-time operation at 100 m/min.	100%	**MAINTAIN** – Median inference latency of 10.5 ms and p95 latency of 11.0 ms support >90 Hz inference, meeting real-time requirements for 100 m/min operation with sufficient spatial resolution for in-line monitoring.
Explainability testing: Support operator oversight through automatic flagging, logging and error review.	95%	**MAINTAIN** – Global and signal-level feature importance aligned with known calendering and ultrasound physics, faithfulness tests passed and importance rankings were stable across runs. Minor limitations stem from dataset size rather than methodological issues.
Consistency and repeatability testing: Support operator oversight through automatic flagging, logging and error review.	20%	**ADDRESS** – Repeated inference on identical inputs produced high output variability (coefficient of variation ≈ 1.0), indicating non-deterministic inference behaviour likely caused by active dropout layers, GPU kernel non-determinism, or floating-point instability.

Benefits realised by testing

- **Risk reduction:** Exposed key failure modes, such as sensitivity to missing UT data and non-deterministic outputs, preventing potential production errors and unreliable measurements
- **Financial impact:** Demonstrated potential to reduce scrap and ramp-up delays, translating to estimated savings of €250 – €350m per gigafactory in the first two years by cutting the scrap from 30% to 10%
- **Informed decision making:** Provided stakeholders with clear performance metrics (e.g., RMSE = 5.74 µm, R^2 = 0.727, >90 Hz inference) to support safe deployment and process optimisation
- **Development pathways:** Identified areas needing improvement, including missing-data handling, OoD detection and repeatability, to guide future model refinement.

Recommendations

- **Improve robustness:** Strengthen handling of missing ultrasound data and ensure stable, repeatable outputs
- **Enhance OoD detection:** Refine out-of-distribution detection and integrate with predictive maintenance to minimise downtime
- **Optimise real-time performance:** Reduce latency and enable adaptive learning for recalibration to changing materials or process conditions
- **Expand explainability:** Introduce operator-centric visualisations that highlight actionable insights rather than raw metrics
- **Collaborate with engineers:** Work closely with process engineers to align model outputs with practical decision-making
- **Integrate into operator interfaces:** Build intuitive interfaces to support scalable deployment and enhance user trust.

Conclusion and recommendations

Trustworthy AI is a prerequisite for safe and successful adoption in UK industry.

As AI capabilities advance at pace, industrial deployment continues to lag – driven not by a lack of ambition, but by a lack of confidence. Independent AI assurance is therefore not optional but an essential infrastructure for unlocking the productivity, sustainability and resilience gains and competitive advantage the UK urgently needs.

In manufacturing, trustworthy AI cannot be validated through software testing alone, it also requires access to realistic digital environments and physical testbeds that reflect the complexity of production systems. These combined infrastructures are essential for evaluating how AI behaves under real operational conditions, something no purely digital simulation can fully capture.

For AI adopters, these case studies demonstrate why performance alone is not enough. Whether implementing internally developed models or procuring external solutions, businesses must be confident that AI systems are safe, reliable, fair, governed and robust under real operating conditions. The HVM Catapult and its strategic partnerships offer the expertise and representative environments needed to validate AI at production scale and help organisations adopt AI safely and responsibly.

For AI assurance providers, this work highlights the complementary value HVM Catapult can bring: industrially relevant cyber-physical testbeds to support awareness, co-creation and demonstration, realistic industrial data, and deep domain expertise that support more rigorous, end-to-end assurance activities than any single organisation can deliver alone.

What is next?

We invite adopters, developers and assurance providers to engage with the HVM Catapult teams, explore how these capabilities can support your need, and collaborate with us in building a safer, stronger and more confident AI ecosystem for UK industry.

This is not only a technical necessity, it is vital to ensure the UK remains at the forefront of responsible, high-value industrial innovation.

Acknowledgements

- Nandini Chakravorti
- Jodie Clark
- Christopher Dungey
- Kate Gongadze
- Nima Hojat
- Sarini Jayasinghe
- Obatarhe Mowoe
- Paulina Mozejko
- Tim Newman
- Mona Faraji Niri
- Dylan Parker
- Narcisa Pinzariu
- Yazan Qarout
- Mostafizur Rahman
- Edwin Anarcaya Roca
- Katharina Roettger
- David Russell
- Hamidreza Farhadi Tolie
- Nick Watson
- Oliver Willis


University of Sheffield
AMRC
Advanced Manufacturing Research Centre

cpi

mtc
where progress happens

ncc
innovating for industry

NMIS
National Manufacturing Institute Scotland

WMG
THE UNIVERSITY OF WARWICK