Appendix C: Workshop Case Study Responses

In this guide

In this guide

Case Study 1: AI-Driven Safety and Regulatory Compliance Evaluation for Manufactured Foods

Question 1. How can the FSA be assured that AI systems used for allergenicity and compositional risk assessments have accessed, interpreted, and applied the correct scientific and regulatory data across all relevant domains?

“AI is a tool for helping-you need to know what questions to ask it.”

“There needs to be some verification that what AI is providing is correct.”

“AI gives you an answer and lulls into false sense of security. People do not give enough thought to the answer.”

AI should be appraised in the same way as any other tool used in food safety assurance. The same safeguards and assurance processes should apply as with non-AI systems. The overall process would not fundamentally change even if AI were deployed widely. Measurement of effectiveness would be through appropriate output and outcome measures. Any negative changes picked up during inspections should be raised with the FBO who should be in a position to provide a satisfactory explanation. If not, it could be a red flag for further investigation.

While processes may not need to change significantly, data quality, accessibility, and sharing must be addressed. Also, AI could support oversight and insights not possible within the current process so processes and procedures might change incrementally in response to opportunities. Opportunities may present that allow for better separation of purely food safety issues versus broader food quality parameters that are not of primary concern to FSA. It was emphasized in the Plenary Session that current UK food regulations are robust and do not require major revision. A complementary perspective was the suggestion that it would be a missed opportunity to limit AI to performing current tasks.

Automation of routine checks is potentially a major opportunity. This might free human agents to devote time to more valuable tasks such as investigating defects or determining root causes of failures.

In well-run food businesses, the addition of AI tools can support and potentially improve food safety controls and assurance. There is a danger in some of the less well-run operations that adoption of AI solutions has the potential to hide poor management by, for example, generating documents and plans that look very good but which are not fully implemented or understood in the organization.

Question 2. What standards should govern the transparency, traceability, and reproducibility of AI-derived risk assessments, particularly when used to justify labelling and safety decisions?

“If had recall – the authority comes in and asks what detection systems do you have in place. I explain to the authorities I use an AI trained x-ray system. The FSA will want to know how the system has been trained.”

“Transparency and traceability is OK. Reproducibility… this is the AI business, not the FSA’s responsibility.”

There was a lack of consensus as to whether new standards are necessary governing, inter alia, transparency, traceability and reproducibility of AI derived risk assessments. If standards were deemed to be necessary, consideration would need to be given to timing, as tools and applications are evolving. The introduction of unnecessary additional standards may well hamper innovation. While standards might be a consideration for future development, current use of AI in risk assessments should be date-stamped as AI will evolve over time.

Participants consistently stressed the importance of maintaining a 'human-in-the-loop' approach. Similarly, AI was seen as a supporting tool (e.g. a co-worker) but not replacing human oversight in critical decisions.

Data are key. More data sharing will be needed. Implementation of AI tools would be supported by high quality information in digital form. Digital transformation of processes is needed. Both of these points have broad implications: data sharing requires trust among different stakeholders in the food system and, at least in some instances, there is a lack of trust. Digital data may require investment and adoption of new technologies. Both changes need concerted action; there are examples where opportunities for digital data sharing were undermined by stakeholders still using legacy systems.

Data governance is needed; data integrity was raised several times. Low input data quality has the potential to lead to poor results and decisions and FBOs should be aware of this.

FSA could support by providing guidance to businesses on the deployment of AI tools. Participants highlighted the opportunity for training; however, training of FBOs would not be for FSA to do.

The term “compliance skill” was used to describe the need of FBOs to understand the food system. In evaluating the significance of any process changes, an understanding of the basis of current controls is important. Microbusinesses and SMEs were highlighted as needing tailored support due to limited resources and technical expertise.

There may be a need to provide training to food officials to ensure awareness of the tools and to ensure they understand the implications. Introduction of AI tools

may generate more complexity from a governance perspective.

Question 3. How do we validate that AI systems can identify emerging risks or uncommon ingredient interactions, and not just replicate existing knowledge—especially when considering public health risk?

“AI can be used to test the risk and the impact of the risk and then model it.”

Emerging risks identified by AI should be validated using methods consistent with current practice. However, AI could be used to rapidly simulate different scenarios in a way that would be difficult or impossible currently.

Case Study 2: AI-Supported Data Pack Generation for Third-Party Certification and Assurance

This section presents an in-depth analysis of the challenges and opportunities in applying AI to automate and augment data pack generation for third-party food certification and assurance. These packs often consist of mixed-format documentation, scanned reports, sensor outputs, handwritten logs; all of which must be synthesized to inform certification decisions. AI systems can offer real-time assistance, retrospective pattern analysis, and expanded coverage for audits, but only if they are implemented within robust governance structures. Manning et al. (2022) argue that responsible AI must be traceable, explainable, accountable, and accessible. These principles must not only be present at deployment but also evolve with continued operation. AI systems deployed in assurance contexts must remain open to human challenge and reinterpretation throughout their lifecycle.

1. How can we assure the robustness, consistency, and context-awareness of multimodal AI systems operating across diverse data sources?

"AI may take falsified records at face value. No change in pen used between shifts could be a red flag.";

"AI can look across the supply chain… auditors only able to look at a small selection of data."

Participants highlighted the complexity of interpreting mixed-format documentation, especially when human behaviours, workarounds, or anomalies exist. AI tools must go beyond surface extraction to understand context. These tools must not only process multimodal inputs, but also infer operational context;

when and where a data recording was written, how it aligns with sensor data, and who entered it. Workshop participants suggested that AI models should be trained on datasets reflecting the full range of operational variability, including uncommon edge cases. The goal is not just information extraction but evidence synthesis; an ability to build narratives from fragmented sources, really as a human auditor would.

"Systems that only read structured fields miss the nuance of how and why data was recorded. AI should learn to reason, not just match templates."

Participants recommended embedding ‘context windows’ into model architectures, enabling models to interpret temporal and spatial relationships within data sets. By assessing the sequence of inputs and recognizing cross-modal inconsistencies, AI can begin to emulate the human capacity for contextual inference. This would help identify when data appears artificially consistent or repetitive, common potential evidence for fraud, and better support assurance judgments that rely on human intuition.

2. What evidentiary standards must AI meet to ensure that its answers to assurance questions are auditable, transparent, and aligned with regulatory interpretation? How do these compare with current practices for human inspectors?

"Two auditors may have a different conclusion both looking at the same factory."

"Two auditors may reach different conclusions; a model must be transparent enough to support its claims with evidence."

The workshop reflected uncertainty about the evidence base required to support AI-generated conclusions. AI must codify and consistently apply regulatory standards. Where interpretation varies by scenario, AI should defer to humans or be configured with confidence thresholds. Explainability must be built into assurance tooling, allowing users to interrogate not only conclusions but also their evidentiary base. Proposed measures included visual traceability maps, structured logic chains, and metadata summaries for each judgement.

"If a human auditor made a claim, they’d have to back it up with evidence. We should expect the same of an AI model."

Participants called for an agreed validation standard that would guide how AI models present audit evidence. This might involve linking each judgement to its data source and identifying assumptions the model made during inference. By

mirroring human evidence standards, AI systems could provide equivalent or superior levels of audit traceability. Some advocated for regulatory ‘assurance blueprints’ that define which decisions must remain under human control.

3. How do we ensure that AI outputs can be validated, challenged, or corrected by human users—without undermining trust or introducing new risks?

"A transparent system builds confidence, built by competent people, not just industry or regulators."

Participants emphasized the importance of embedded challenge mechanisms and proportional validation frameworks. AI interfaces should allow users to flag unexpected outputs, request supporting evidence, and provide corrective feedback. This feedback must be traceable, auditable, and, where appropriate, incorporated into model refinement. Suggestions included escalation paths, feedback loops for retraining, and role-based permissions for challenging AI conclusions.

"It’s not just having a ‘challenge button’, the system should also log how often and by whom it’s used."

A system’s ability to maintain trust depends on its capacity to support structured disagreement. Attendees proposed audit logs of all user interactions, including overrides and challenges, to track patterns and flag areas of recurring concern. Such logs could inform policy refinement and potentially identify systematic weaknesses in model performance or user interpretation.

4. How does such a system adapt to new regulations and standards over time, which are sometimes rapidly changing?

"If you want AI answers to be the truth, you need to rebuild and retrain models after every regulatory{standards} update."

Participants underscored the need for model agility and standards integration. Proposals included the release of machine-readable updates by standards bodies, co-designing rules with developers, and maintaining version-controlled models linked to evolving standards. Standards sandboxes were suggested for testing tools under future state conditions without disrupting assurance workflows.

"The challenge isn’t just that standards change, it’s that interpretations change too."

Attendees proposed collaborative working groups to interpret changes in practice and provide pre-configured rule updates to developers. Others supported embedding legal logic frameworks in assurance platforms that alert users when changes in standards affect data interpretation or system confidence. In fast-moving standards environments, continuous professional development will be essential, not only for AI systems, but for human overseers as well.

General Reflections and Conclusion

The use of AI in third-party certification presents substantial opportunities for scaling audit capacity and enhancing traceability. However, systems must be grounded in principles of transparency, accountability, and human oversight. Throughout the workshop, participants consistently emphasized the importance of hybrid decision-making models where AI supports rather than replaces human inspectors.

A key takeaway was the importance of building trust not only in the system's technical performance but also in its governance and resilience. Participants warned against over-reliance on AI in contexts where the human touch/ auditor skill remains critical.

Going forward, effective deployment will require close collaboration between system designers, certifiers, standards bodies and end users. Robust training, version control, explainable outputs, and consistent human-in-the-loop validation processes will be critical to ensuring AI supports food assurance safely and effectively.

Participants also emphasised the scale advantage AI offers. Unlike traditional auditors who sample selectively, AI systems can scan entire datasets across supply chains in real-time, identifying continuous trends and anomalies with greater precision. This capacity can support auditors by flagging areas requiring human interpretation. However, the utility of such wide-ranging surveillance is conditional upon maintaining data quality and interpretability across heterogeneous systems.

There was concern that human audit consistency varies significantly. AI could bring a more standardised evaluation baseline, but only if benchmarked effectively against multiple auditors and real-case discrepancies. Transparency in how AI reaches conclusions was considered critical, particularly where automated systems supplement or replace human decision-making.

Several contributors underscored that human audits are not infallible, and the benchmark for AI should be parity with, or improvement on, current outcomes. The emphasis was not on perfection, but on traceability and the availability of mechanisms to flag, interrogate, and override AI errors. Ensuring that the supporting evidence remains auditable was seen as key to maintaining public and regulatory trust.

Experts warned that systems adapting to shifting standards and regulations must undergo rigorous revalidation. Quick updates, while technically feasible, may lead to untrustworthy outputs if not properly verified. Establishing protocols for AI lifecycle management, including regular re-training and validation cycles, was considered essential to sustaining compliance over time.

Case Study 3: AI-Enabled Pathology Detection in Abattoirs

This section presents an analysis of the challenges, opportunities, and regulatory implications of integrating AI systems into meat inspection processes at UK abattoirs. Based on findings from the FSA workshop and grounded in responsible AI principles (Manning et al., 2022), the discussion focuses on the use of AI for real-time detection of pre- and post-mortem pathologies. The section addresses critical regulatory questions relating to accuracy, data quality, validation, and governance frameworks, while emphasising the importance of human oversight in high-risk environments. Manning et al. (2022) highlight that responsible AI must be designed to be traceable, explainable, accountable, and accessible, with these properties evolving over time as systems are deployed and interact with complex environments. They argue that AI must remain open to human challenge and reinterpretation, not only at design but throughout its lifecycle, reinforcing the need for transparency and continuous learning.

How can regulators ensure that AI systems for pathology achieve, and maintain, accuracy, recall, and stability at a level equivalent to or exceeding that of trained human inspectors?

"Systems need to be tested over a long enough period to cover rare pathologies and ensure performance in operational settings." Ensuring AI systems match or surpass human performance requires validation under real-world conditions. During the workshop, participants emphasised the need for longitudinal trials, ideally spanning six months or more, during which AI decisions are shadowed and reviewed by human inspectors. Such trials must encompass a diverse range of carcasses and environmental conditions to simulate operational variability.

"Performance should be monitored not just at launch but continually, systems need metrics tied to human oversight." AI model performance should be benchmarked against human inspection rates for both true positives and false negatives. Importantly, the system should provide confidence scores and justifications for its outputs, thereby supporting transparency and enabling human reviewers to make informed decisions. Participants supported a ‘closed-loop’ assurance cycle, in which real-world outcomes inform model retraining. In terms of standards, routine recalibration and scenario testing would be mandated as part of system governance, especially when model performance falls outside acceptable thresholds. Continuous monitoring, supported by a secure audit trail would ensures that AI does not degrade over time or drift away from its validated state.

What standards and validation processes should be established to evaluate the diversity and quality of training data, especially for rare or emergent pathogens where symptoms may not be well represented in existing datasets?

“There is a need for richer training data, particularly around rare or emergent conditions, with variation in lighting, angle, and pathology type”. Training data quality was highlighted as a critical foundation for AI success. Diverse, high-resolution imagery from multiple processing sites is required to reflect the full range of carcass conditions, including lighting variability, camera angles, and rare pathology types. Participants emphasised that without rich training datasets, AI tools could fail to detect uncommon or emerging conditions, thereby compromising food safety. Workshop recommendations included establishing a centralised and anonymised dataset repository. This would allow for consistent evaluation of AI tools and potentially facilitate data sharing by technology developers.

"Models should be tested against diverse and anomalous datasets to validate robustness in the real world." Annotated datasets should include metadata on inspection context, species, and confirmed pathology outcomes to support meaningful benchmarking. Validation should include stress-testing models with deliberately ambiguous or noisy inputs to assess robustness. Standards should mandate the inclusion of under-represented conditions and specify acceptable minimum data volumes for new model releases. The aim is to ensure representativeness, fairness, and performance consistency across use contexts. Models should be tested against diverse (including anomalous) datasets to validate robustness in the real world.

What criteria must be met before AI systems can be authorised for use in regulated environments, and how can ongoing performance be assured, particularly in terms of drift, bias, or unforeseen failure?

"We need to monitor for model drift, bias and maintain performance, unexpected failures need detection mechanisms." Authorisation of AI systems in regulated environments requires the development of a formal certification framework. Such a standards framework should define pre-market testing conditions, performance thresholds, data traceability requirements, and human-AI interface standards. Workshop participants proposed a tiered deployment model beginning with supervised trials, followed by staged rollout with increasing autonomy. Each deployment stage would require predefined success metrics, potentially with regulatory oversight at every phase.

"There should be independent checks for fairness, especially across sites using different camera technologies or lighting setups." To guard against bias and drift, systems must incorporate self-monitoring tools and notify operators if key indicators deviate from expected norms. Manning et al. (2022) suggest lifecycle governance approaches, including automated alerts and scheduled revalidation. Independent auditors could be employed to conduct performance and bias assessments at regular intervals. Any material changes to AI architecture or training data should trigger re-validation/certification. There should be independent checks for fairness, especially across sites using different animals, batches, camera technologies or lighting setups.

General Reflections and Conclusion

"We must define where human judgement must override AI, particularly for ambiguous or edge cases." Beyond technical performance, the ethical dimensions of AI use in abattoirs were a recurring concern. Participants stressed that AI should not become a substitute for human judgement in food safety. The ‘two in a box’ model, wherein AI works in tandem with a human inspector, emerged as a governance mechanism. This model balances efficiency with oversight, ensuring that AI augments rather than replaces human expertise. Participants also warned of ‘automation creep’, where AI systems gradually assume decision-making roles without explicit governance sanction. To mitigate this, policy should consider clearly defined thresholds beyond which only a human can make or validate decisions. Furthermore, retraining and upskilling programmes are vital to ensure inspectors remain capable of effective oversight in increasingly digital inspection environments.

The integration of AI into abattoir inspections promises substantial gains in safety, efficiency, and consistency. However, realising these benefits safely depends on transparent governance, robust validation, ethical safeguards, and a commitment to sustaining human expertise. AI must serve as a co-pilot, not a replacement, within the well-defined and understood regulatory frameworks. We must define where human judgement must override AI, particularly for ambiguous or edge cases.

Case Study 4 – AI-Powered Document Inspection at UK Ports of Entry

Background

The UK Food Standards Agency along with software developers are exploring the use of AI systems, including large language models (LLMs), to enhance the inspection and verification of food import documentation at ports. These documents, ranging from health certificates and commercial invoices to packing lists and shipping manifests, are critical for ensuring food safety, regulatory compliance, and traceability of goods entering the UK. Traditionally, official controls involve officers manually reviewing these documents to assess conformity with safety standards and detect inconsistencies or fraudulent entries. This process can be time-consuming, especially under increased trade volumes and complex global supply chains.

An AI-based solution, incorporating document classification models, optical character recognition (OCR), and LLMs, could increase the productivity of frontline officers (e.g., freeing up time for more physical inspections and investigations). These systems can automatically extract key data points, cross-check documents for internal consistency, flag anomalies or incomplete submissions, and even interpret unstructured or multilingual content. LLMs, specifically, have shown promise in identifying subtle discrepancies in language, such as ambiguous product descriptions or suspicious edits.

1. How can the accuracy, reliability, and auditability of LLMs be assured when used to assess official import documentation in regulated environments?

“LLMs are helpful but only partly, looking to narrow it down for greater degree of success.”
“Need to monitor ongoing performance of the system. People are always looking for new ways to bypass systems.”
“Using humans in the loop to improve.”

“AI can look for spikes and anomalies that flag the need for further investigation.”

LLMs are a new ‘cultural technology’ that allows individuals to take advantage of collective knowledge, skills and information accumulated through human history. Implicit elements of intelligence that are largely omitted from current models include: wisdom and judgement developed through experience; creative insight that transcends pattern recombination; intuitive understanding that cannot be verbalized; embodied knowledge learned through physical interaction; and self-awareness and metacognition. Many of these omissions featured in the workshop discussion.

The current priority at ports is food safety but if OCR and LLMs can remove some of the burden of risk assessment from inspectors then more time could be spent on ensuring authenticity, and preventing fraud and smuggling. The purpose/question underpinning document inspection needs to be clear and judgements about acceptability of the paperwork need to be consistent. The system needs to cope with documents from a variety of sources including printed, handwritten and scanned material of low resolution. Assessments vary in complexity. In addition to registering right/wrong responses to questions, the system could also look for anomalies (e.g. a spike of a product entering from a region/country not seen before) requiring further investigation. The ongoing performance of the system needs to be monitored and improved in the light of experience.

2. What safeguards are required to manage the risk of false positives or negatives, omissions, or AI-generated hallucinations, especially when decisions impact food safety or border clearance?

“It’s not just having a ‘challenge button’, the system should also log how often and by whom it’s used.”
“We still need humans to sample even when it looks like it’s working.”

“Confidence levels should be known”
“Explainability must be built into assurance”

Current AI systems are designed to produce an answer with hallucination occurring when the system generates plausible but incorrect information. Continuous checking, retraining and validation are necessary to mitigate these risks.

A potential benefit of AI systems is that an audit approach could be employed giving a feel for intent and corporate activity of the organization. While some decisions are binary, others are more nuanced with human judgement required. Confidence levels and contextual analysis could be part of the AI output to inform human decisions.

The system should be gradually introduced with clear guidance on the roles of machines and humans in decision making. A continuous learning approach should be adopted with the role of humans changing as AI systems learn. AI systems can be used to detect anomalies leading to human interventions to mitigate errors in decision-making.

3. How to validate training data diversity and alignment with UK regulatory terminology, languages, and document formats to ensure equitable and robust performance?

“You never get perfect data. You get imperfect data and work with what you have.”
“These systems must align with our forms, terms and regulation and be updated when regulations change.”
“If you want AI answers to be the truth, you need to rebuild and retrain models after every standards update.”

Participants stressed that AI models used at UK ports must be trained on datasets that reflect the full diversity of documentation encountered, including non-English content, handwritten submissions, and forms generated under differing regulatory regimes. Without this diversity, models may perform well in ideal cases but fail in realistic or marginal ones.

Alignment with UK-specific terminology and standards was seen as essential. This includes updating models when guidance changes, and fine-tuning LLMs on validated local data. Validation should therefore be continuous. In practice, this may include shadow testing (comparing AI outputs against human decisions), spot audits, and real-time confidence scoring. Ultimately, data diversity and regulatory alignment should be treated as core governance requirements rather than optional refinements.

4. What level of accuracy by an AI system is considered acceptable and would result in reducing burden on human inspectors? Should an AI system result be binary (pass or not) or should it provide a more nuanced output that includes reasons for an assessment?

“AI should be nuanced and leave the decision up to the human.”
“You need a reason why something isn’t allowed in.”

Participants agreed that no fixed accuracy threshold would universally justify removing human oversight. Rather, acceptable performance must be judged in context, by comparing AI output to current inspection accuracy and by determining how the tool complements rather than replaces human judgement.

Binary outputs may be useful for some high-confidence cases but were generally seen as insufficient for complex scenarios. A tiered system was suggested, where AI models provide a confidence score and rationale, enabling human inspectors to decide whether further investigation is needed. Such systems would allow inspectors to prioritize workloads more effectively, directing attention to ambiguous or borderline cases.

Ultimately, AI should enable better decisions, not faster errors. Participants emphasized that clarity of explanation, traceability of decision logic, and ongoing human review would be key determinants of a model’s fitness for purpose.