

ORIGINAL RESEARCH

Open access

# Explainable AI for Food Safety Decisions: A Framework for Risk Interpretation, Regulatory Confidence, and Operational Accountability

Lars Andersen<sup>1\*</sup>, Freya Nielsen<sup>1</sup>, Mette Sørensen<sup>2</sup>

## Abstract

Artificial intelligence is increasingly used in food safety systems to support contamination detection, shelf-life prediction, quality classification, inspection prioritisation, and risk scoring. These applications promise faster, more sensitive, and more scalable decision support across complex food supply chains. However, the same predictive complexity that makes AI useful can also make its outputs difficult for food safety professionals, regulators, and operators to interpret. The central problem addressed in this framework article is that AI-supported food safety decisions may become opaque, uncontestable, and weakly aligned with existing structures of regulatory responsibility. When risk scores or alerts cannot be understood, challenged, or traced to relevant evidence, they may undermine rather than strengthen food safety governance. This problem is especially serious because food safety decisions involve direct public health consequences, market consequences, and legal accountability. The objective of this article is to develop an original Explainable AI Framework tailored to food safety decision systems. The framework integrates four interdependent dimensions: risk interpretation, regulatory confidence, operational accountability, and human review. It treats explainability not only as a technical property of models but also as a governance condition for responsible food safety decision-making. The article is based on a conceptual synthesis of 27 peer-reviewed articles published between 2017 and 2025 across food safety, explainable artificial intelligence, algorithmic accountability, risk assessment, and digital food safety systems. It does not present new empirical data. Instead, it constructs a design and governance framework that can guide the development, validation, deployment, and oversight of AI-supported food safety tools. The proposed framework defines four pillars: interpretable risk communication, transparent regulatory evidence, accountable operational responsibility, and contestable human review. Together, these pillars show how food safety AI systems can become not only predictive but also auditable, trustworthy, and practically usable. The article argues that embedding explainability into food safety AI is a governance and trust-building imperative rather than a purely technical enhancement.

**Keywords** Food safety, Explainable AI, Framework, Risk interpretation, Regulatory confidence, Operational accountability

\*Correspondence:

Lars Andersen  
laa@agro.au.dk

<sup>1</sup> Department of Agro-Environmental Engineering, Faculty of Technical Sciences, Aarhus University, Aarhus, Denmark

<sup>2</sup> Department of Intelligent Food Production, Faculty of Science, University of Copenhagen, Copenhagen, Denmark

## Introduction

Artificial intelligence is becoming an important component of modern food safety systems because it can process large, heterogeneous, and time-sensitive data streams more rapidly than conventional manual review. Machine learning has

been applied to food safety monitoring, early warning, contamination prediction, quality classification, and risk prioritisation, while broader digitalisation is reshaping how food systems detect and respond to hazards [1]. The growth of big data, Internet of Things infrastructure, and predictive analytics has expanded the capacity of food safety organisations to anticipate risks rather than only react after incidents occur [2, 3]. Yet this expansion creates a new governance problem: the more complex AI systems become, the harder it may be for decision-makers to understand why a specific alert, score, or classification was produced.

Food safety is not a low-stakes optimisation environment because decisions may affect consumer health, regulatory enforcement, product recalls, trade, brand reputation, and legal responsibility. Reviews of AI and machine learning in food safety emphasise that these technologies can improve monitoring and prediction, but they also require careful validation and interpretation before their outputs can be incorporated into operational decisions [4, 5]. In this context, opacity is not merely a usability problem; it can weaken the chain of evidence linking data, model output, managerial judgement, and regulatory action. Explainable AI therefore becomes necessary when predictive systems influence decisions that must be justified to inspectors, auditors, customers, and public authorities.

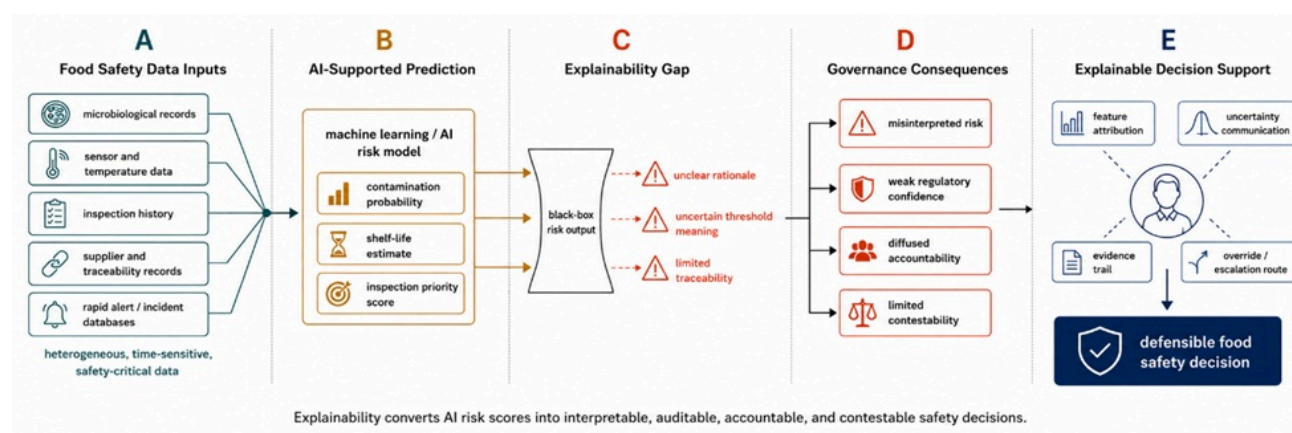
The wider XAI literature shows that explainability is commonly framed as a response to the limitations of black-box prediction in high-stakes settings. Taxonomies of XAI distinguish between transparency, post hoc explanation, interpretability, fidelity, and user-centred intelligibility, all of which matter when AI outputs must support accountable decisions [6, 7]. Rudin argues that black-box systems should not be normalised in high-stakes contexts when interpretable alternatives can provide decision support with clearer reasoning [8]. For food safety, this argument is particularly important because the acceptability of a model depends not only on statistical performance but also on whether its reasoning can be inspected, challenged, and connected to domain knowledge.

This article develops an original Explainable AI Framework for food safety decisions by synthesising food safety AI research with scholarship on explainability, accountability, trust, and contestability. It focuses on four linked problems: how risk is interpreted, how evidence becomes transparent to regulators, how operational responsibility is allocated, and how human experts can review or contest AI-generated recommendations. The framework is conceptual and governance-oriented rather than empirical, but it is grounded in existing work on machine learning applications in food safety and broader debates on accountable AI [9, 10]. The article proceeds by explaining why food safety is an explainability challenge, analysing risk interpretation, and then developing the framework's regulatory, operational, and human review logic.

## Food Safety as an Explainability Challenge

Food safety creates distinct explainability demands because hazards emerge from complex interactions among biological, chemical, physical, environmental, and operational factors. Predictive systems may draw on microbiological records, supply-chain alerts, inspection histories, spectral data, environmental monitoring, or rapid alert databases, but these data sources do not automatically translate into meaningful safety judgements. Research on machine learning in food safety monitoring shows that predictive models can support incident detection and risk forecasting, yet their usefulness depends on whether food safety professionals can understand the basis of predictions [4]. In this setting, explainability must connect model behaviour to hazard pathways, control points, product categories, and operational decisions.

**Figure 1** illustrates how AI-supported food safety decisions become explainability challenges when predictive outputs are not connected to risk interpretation, regulatory evidence, operational responsibility, and human review.



**Figure 1.** Explainability Challenges in AI-Supported Food Safety Decisions: From Heterogeneous Safety Data to Risk Outputs, Governance Gaps, and Human Review Needs

The complexity of food matrices and contamination pathways means that generic AI explanations may be insufficient for food safety use. A feature importance score can identify variables associated with risk, but it may not explain whether the risk is microbiologically plausible, operationally controllable, or relevant to a specific process step. Reviews of AI in food safety highlight the diversity of applications, from contamination detection to risk assessment and decision support, which implies that explanation needs differ across use cases [1, 5]. Therefore, food safety explainability must be domain-specific rather than limited to generic visualisations or statistical summaries.

Another challenge is that food safety decisions are embedded in regulatory and legal structures that require traceability, documentation, and justification. Digital food safety systems and early warning tools are increasingly expected to support preventive risk management, but such systems can only strengthen governance if their outputs are linked to transparent evidence and reviewable decision pathways [11]. When AI systems inform recalls, inspection targeting, or supplier risk classification, the organisation must be able to explain why a decision was made and who authorised it. This creates a direct connection between explainability, regulatory confidence, and operational accountability.

Food safety AI also raises trust calibration challenges because operators may either over-rely on algorithmic alerts or dismiss them when they conflict with experience. General XAI research shows that explanations can improve user understanding, but they can also create misplaced confidence if they are incomplete, misleading, or poorly matched to user expertise [12, 13]. In food safety, this means that explanations must be designed for specific decision roles, including quality assurance staff, plant managers, inspectors, risk assessors, and regulators. The goal is not simply to make the model visible, but to make its reasoning usable for safe, proportionate, and defensible action.

## Risk Interpretation in AI-Supported Decisions

AI-supported food safety decisions often take the form of risk scores, classification labels, anomaly alerts, probability estimates, or prioritisation rankings. These outputs can appear precise, but their operational meaning depends on how users interpret uncertainty, thresholds, and consequences. Studies using rapid alert and food safety databases show that AI can help predict or classify safety risks, yet a score alone does not determine whether an inspection, recall, hold decision, or additional test is warranted [14, 15]. Risk interpretation therefore requires a translation layer between algorithmic output and practical food safety judgement.

The interpretation problem is partly cognitive because users must assess whether an AI output is credible, actionable, and aligned with known hazard mechanisms. XAI methods can support this process by indicating which variables influenced a prediction, but explanations must be understandable to the intended audience and faithful to the model's behaviour [6, 7]. For example, feature attribution may help a food safety manager identify whether a high-risk

classification was driven by product category, origin, prior alert history, storage condition, or microbiological signal. However, the explanation is useful only if it helps the user judge whether the alert reflects a real hazard, a data artefact, or a need for further evidence.

Risk interpretation is also shaped by trust calibration, because users must neither accept nor reject AI outputs automatically. A highly complex model may produce accurate predictions but still fail operationally if staff cannot understand when to rely on it or when to escalate for human review. Work on trustworthy AI emphasises that reliability, transparency, fairness, accountability, and human-centred design must be considered together rather than treated as separate technical objectives [13]. In food safety, calibrated trust means that AI should support expert judgement without displacing professional responsibility for risk-based decisions.

Different food safety applications create different explanation requirements because the decision context, data type, and consequences vary. A shelf-life prediction model may require explanations about temperature history and product characteristics, while a fraud detection tool may require evidence about unusual supply-chain patterns or analytical signatures. Food fraud detection using explainable AI and microbiome-based origin classification show that explanation techniques can make complex food-related predictions more interpretable when linked to domain-relevant evidence [16, 17]. **Table 1** outlines the key dimensions of risk interpretation and the explainability requirements for different AI-based food safety applications.

**Table 1.** Risk Interpretation in AI-Supported Food Safety Decisions: Application Types, Explainability Needs, and Human Factors

| AI-based food safety application    | Typical AI output                                                         | Core risk interpretation challenge                                                                                | Explainability requirement                                                                                                          | Human factor requiring attention                                                                                  |
|-------------------------------------|---------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------|
| Contamination detection             | Probability of contamination, anomaly flag, or hazard class               | Translating a statistical signal into a decision about product hold, testing, release, or recall                  | Clear evidence showing which analytical, microbiological, environmental, or process variables contributed to the alert              | Avoiding over-reaction to false positives and under-reaction to plausible low-frequency hazards                   |
| Shelf-life and spoilage prediction  | Remaining shelf-life estimate, quality class, or deterioration risk score | Understanding whether the prediction reflects product condition, storage history, packaging, or model uncertainty | Explanation of time-temperature effects, product attributes, and uncertainty around the predicted shelf-life window                 | Supporting proportionate judgement by quality assurance staff rather than automatic acceptance of model estimates |
| Inspection prioritisation           | Risk ranking for facilities, suppliers, products, or batches              | Converting a ranking into transparent inspection priorities that can be justified to managers and regulators      | Traceable ranking logic showing the contribution of prior incidents, product type, supplier history, and contextual risk indicators | Preventing blind trust in rankings and ensuring that expert knowledge can modify priorities when justified        |
| Food fraud or origin authentication | Fraud probability, authenticity classification, or origin assignment      | Distinguishing genuine fraud signals from natural variation in product composition,                               | Domain-specific feature attribution linking model output to meaningful compositional,                                               | Ensuring that explanations are understandable to technical experts and                                            |

|                                           |                                                      | geography, or processing                                                        | microbiome, spectral, or supply-chain indicators                                                                         | legally defensible for enforcement decisions                                                                |
|-------------------------------------------|------------------------------------------------------|---------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------|
| Early warning and emerging risk detection | Emerging hazard alert, cluster signal, or risk trend | Assessing whether weak signals are credible enough to justify preventive action | Transparent signal provenance, confidence level, trend explanation, and connection to known or plausible hazard pathways | Balancing precaution with alert fatigue and ensuring that operators can challenge weak or ambiguous signals |

## Regulatory Confidence and Evidence Transparency

Regulatory confidence in AI-supported food safety decisions depends on whether model outputs can be connected to evidence that is valid, auditable, and relevant to recognised safety obligations. Food safety authorities and auditors do not evaluate decisions only by their predicted accuracy; they also require documentation of data provenance, validation logic, uncertainty, and the rationale for action. The XAI literature shows that explanation is most valuable when it supports inspection, verification, and institutional trust rather than merely producing simplified visual outputs [18]. In food safety, this means that an AI alert must be supported by a transparent evidentiary trail that can be reviewed before it becomes the basis for regulatory or operational intervention.

Evidence transparency is especially important when AI systems are built on complex data sources such as rapid alert databases, sensor streams, inspection histories, and supply-chain records. Transformer-based approaches to food safety risk prediction using RASFF data illustrate how advanced models can identify patterns across large alert systems, but their regulatory usefulness depends on whether their reasoning can be explained in terms of recognisable hazard, product, and origin relationships [19]. Risk classification matrices for food incidents provide one way to translate AI-supported outputs into structured categories that can be understood by food safety professionals and regulators [20]. Such tools help bridge the gap between statistical prediction and regulatory judgement.

The alignment between AI and preventive control systems also requires that model outputs be mapped onto existing food safety processes. HACCP monitoring and animal-source food safety applications show that AI can support preventive control, but only if alerts are connected to hazards, control points, corrective actions, and verification steps [21]. Digitalisation and AI may contribute to more sustainable and responsive food systems when they strengthen rather than bypass established food safety responsibilities [22]. Therefore, regulatory confidence depends on embedding explainability into the operational language of hazard analysis, preventive monitoring, and documented corrective action.

Regulatory confidence also requires that AI systems remain inspectable after deployment because models may drift as products, suppliers, processes, and hazards change. Early warning tools for emerging risks demonstrate the value of digitally driven detection, but they also show that evidence must remain traceable when weak signals are converted into food safety action [23]. The problem of inscrutability is not solved by declaring that a model is accurate; it is addressed by making the model's inputs, assumptions, outputs, uncertainty, and review history available for scrutiny [24]. A regulatory-ready AI system must therefore provide explanation as part of a continuing evidence lifecycle rather than as a static technical appendix.

## Operational Accountability in Food Safety Systems

Operational accountability concerns the allocation of responsibility when AI systems influence food safety decisions. In conventional food safety systems, responsibility is assigned through roles such as quality assurance manager, food safety team leader, production supervisor, auditor, or regulator. AI can complicate this structure by distributing influence across software developers, data providers, model validators, system owners, and human operators [9]. Unless

responsibility is explicitly designed into the system, an unsafe decision may become difficult to attribute, explain, or correct.

The risk of diffused accountability is particularly serious when AI tools are used for inspection prioritisation, product release, contamination alerts, or supplier risk classification. Algorithmic accountability research shows that organisations must define who is responsible for model design, deployment, interpretation, override, and post-decision review [10]. In food safety, this means that AI recommendations should not enter operational workflows without clear ownership of the final decision. The framework proposed here treats accountability as a design requirement that must be specified before deployment rather than repaired after failure.

Operational accountability also requires traceability from the AI output to the human decision that follows it. Food safety risk prediction systems based on alert data can support decision-making, but they must record which evidence was displayed, how the risk was interpreted, and whether the human user accepted, modified, escalated, or rejected the recommendation [14, 15]. This traceability is essential for learning from false positives, false negatives, near misses, and corrective actions. It also protects responsible operators by showing that decisions were made through a documented and reviewable process.

A practical accountability structure should therefore distinguish between algorithmic responsibility and decision responsibility. The AI system may be responsible for producing a ranked alert, classification, explanation, or uncertainty estimate, but the organisation must define who has authority to act on that output. Accountable AI scholarship warns that human responsibility can become symbolic when oversight is poorly defined or when users lack the information needed to intervene meaningfully [9, 25]. Food safety organisations should therefore assign explicit decision owners, escalation routes, and review duties for every AI-supported safety action.

## Human Review and Decision Contestability

Human review is necessary because food safety AI systems operate in contexts where data may be incomplete, noisy, biased, delayed, or inconsistent with local process knowledge. Even highly capable models cannot fully replace expert judgement about product history, plant conditions, supplier reliability, sanitation deviations, or unusual operational events. Contestability research argues that affected users and responsible decision-makers must be able to question algorithmic decisions rather than merely receive them as fixed outputs [26]. In food safety, contestability means that operators, quality staff, and regulators should be able to ask why an alert occurred and whether it should be accepted.

Human review should not be reduced to a superficial approval button because that creates the appearance of oversight without meaningful control. Green's critique of human oversight policies shows that oversight may fail when humans lack time, authority, expertise, or sufficient explanation to challenge algorithmic outputs [25]. For food safety, this means that review systems must provide actionable information, not just model scores. A human reviewer should see the evidence behind the recommendation, the level of uncertainty, the relevant control point, the recommended action, and the consequences of accepting or rejecting the alert.

Contestability also requires second-opinion mechanisms that allow uncertain or high-impact AI outputs to be escalated. For example, an AI-generated contamination alert may require confirmation through additional testing, expert review, supplier inquiry, or inspection before a costly recall or product hold decision is made. Explainable AI tools can support this escalation by showing which variables or patterns drove the recommendation and whether similar cases previously resulted in confirmed hazards [6, 13]. The goal is to ensure that AI accelerates expert reasoning without closing off deliberation.

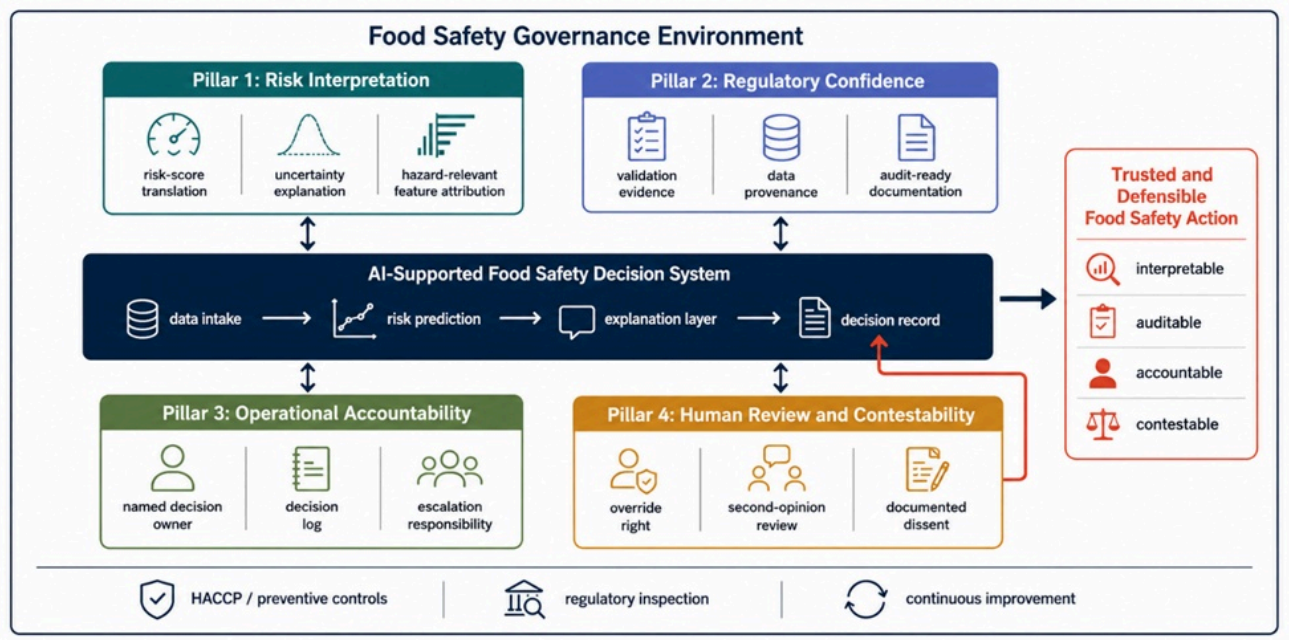
A contestable food safety AI system should also preserve the right to override, document dissent, and learn from disagreement. When an experienced operator challenges an AI recommendation, the system should record the reason for disagreement and feed that information into model monitoring and governance review. This design approach is

consistent with the broader argument that explanations should be tailored to human purposes, including understanding, action, correction, and accountability [7, 12]. In food safety, contestability transforms explanation from a passive description into an active mechanism for safer decision-making.

# Proposed Explainable AI Framework

The proposed Explainable AI Framework is built on four interconnected pillars: risk interpretation, regulatory confidence, operational accountability, and human review. Risk interpretation ensures that AI outputs are translated into meaningful food safety judgements; regulatory confidence ensures that those judgements are supported by transparent and auditable evidence. Operational accountability defines who owns the decision and how responsibility is documented, while human review ensures that AI recommendations remain contestable and subject to domain expertise. Together, these pillars convert AI from a black-box predictor into a governed decision-support system [6, 9].

Figure 2 presents the proposed Explainable AI Framework as an integrated governance cycle linking risk interpretation, regulatory confidence, operational accountability, and contestable human review.



**Figure 2.** Proposed Explainable AI Framework for Food Safety Decisions: Four-Pillar Governance Cycle for Interpretable, Auditable, Accountable, and Contestable AI Use

The first pillar, risk interpretation, requires that every AI output be accompanied by an explanation appropriate to its application type, data source, and decision consequence. In contamination detection, this may involve identifying the analytical or process variables most responsible for a hazard alert; in fraud detection, it may involve explaining unusual supply-chain, compositional, or microbiome signals [16, 17]. The second pillar, regulatory confidence, requires evidence transparency through validation records, data provenance, uncertainty reporting, and audit trails. These mechanisms allow regulators and auditors to evaluate whether AI-assisted decisions are consistent with food safety obligations and preventive control logic [20, 21].

The third pillar, operational accountability, requires that food businesses define decision ownership before AI tools are deployed. This includes specifying who reviews alerts, who approves corrective actions, who can override the system, and who is responsible for documenting outcomes. The fourth pillar, human review, requires contestability mechanisms such as override rights, escalation pathways, second-opinion review, and post-decision learning [25, 26]. Table 2

presents the proposed Explainable AI Framework integrating risk interpretation, regulatory confidence, operational accountability, and human review.

**Table 2.** Proposed Explainable AI Framework for Food Safety Decisions: Pillars, Design Principles, and Governance Mechanisms

| Framework pillar           | Core design principle                                                                                  | Required AI explanation mechanism                                                                              | Governance mechanism                                                                                                | Expected food safety contribution                                                                                    |
|----------------------------|--------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------|
| Risk interpretation        | AI outputs must be translated into meaningful safety judgements that users can understand and act upon | Feature attribution, uncertainty communication, threshold explanation, and domain-specific risk narratives     | User training, explanation templates, decision thresholds, and risk communication protocols                         | Improves the ability of quality assurance staff and managers to interpret alerts, probabilities, and classifications |
| Regulatory confidence      | AI-supported decisions must be supported by transparent and inspectable evidence                       | Validation summaries, data provenance records, model performance documentation, and audit trails               | Regulatory documentation, periodic validation, audit access, and alignment with HACCP or preventive control records | Strengthens regulatory acceptance by making AI outputs reviewable, evidence-based, and traceable                     |
| Operational accountability | Responsibility for AI-assisted decisions must be assigned before deployment                            | Decision logs showing model output, explanation viewed, human action, and justification                        | Named decision owners, escalation rules, approval workflows, and responsibility matrices                            | Prevents diffused accountability and clarifies who is responsible for acting on or rejecting AI recommendations      |
| Human review               | Human experts must be able to contest, override, and learn from AI recommendations                     | Counterfactual explanation, comparison with similar cases, confidence warnings, and alternative action prompts | Override rights, second-opinion review, appeal pathways, and post-decision learning loops                           | Ensures that domain expertise remains central and that AI does not close down professional judgement                 |
| Integrated governance      | Technical explanation and organisational responsibility must operate as one system                     | Combined explanation dashboard linking risk, evidence, responsibility, and review status                       | Cross-functional governance board, monitoring schedule, incident review, and continuous improvement cycle           | Produces a transparent, trusted, and accountable AI-supported food safety decision system                            |

The framework should be implemented as a continuous governance cycle rather than a one-time model documentation exercise. AI models must be selected, validated, explained, deployed, monitored, contested, and revised through a process that connects technical performance with food safety responsibility. This approach is consistent with digital food safety research showing that AI, big data, and early warning systems can strengthen resilience when they are embedded

within accountable risk governance [11, 23]. The framework therefore positions explainability as the organising logic that links prediction, evidence, responsibility, and review.

## Practical Implications and Future Research

For food businesses, the framework provides a practical implementation pathway for adopting AI without weakening food safety accountability. Organisations should begin by classifying AI applications according to decision consequence, data type, and regulatory sensitivity, then select explanation techniques that match those use cases. SHAP, LIME, rule extraction, attention visualisation, and interpretable models may each be useful, but their suitability depends on whether they produce explanations that food safety professionals can connect to hazards, control points, and corrective actions [6, 8]. Training should therefore focus not only on how to use an AI tool but also on how to interpret risk, challenge outputs, and document decisions.

For regulators and auditors, the framework suggests that AI assessment should extend beyond model accuracy to include explanation quality, evidence transparency, auditability, and accountability design. Digitalisation in food systems creates opportunities for more proactive and sustainable safety governance, but these opportunities depend on whether AI systems can be inspected and justified in practice [22]. Regulatory engagement should therefore occur early in the AI lifecycle, particularly for tools that affect inspection targeting, product release, supplier evaluation, or recall decisions. This engagement can help align AI design with preventive controls, HACCP records, and evidence expectations before deployment.

Future research should empirically test whether explainability improves food safety decision quality, inspection efficiency, operator trust calibration, and regulatory confidence. Food-specific XAI benchmarks are needed because generic explainability measures may not capture whether an explanation is meaningful for microbiological, chemical, spectral, supply-chain, or fraud-related decisions [1, 5]. Studies should also examine how different user groups respond to AI explanations, including plant-level quality staff, corporate food safety managers, auditors, and regulators. The next phase of research should move from asking whether AI can predict food safety risks to asking whether AI-supported decisions can be understood, contested, and governed responsibly [27].

## Conclusion

This article proposed an Explainable AI Framework for food safety decisions that integrates risk interpretation, regulatory confidence, operational accountability, and human review. The framework responds to the growing gap between the predictive capabilities of AI systems and the governance mechanisms needed to make their outputs transparent, contestable, and responsible. Its central contribution is to show that explainability in food safety is not only a model property but a decision-system requirement.

The framework argues that food safety AI should not be evaluated only by predictive accuracy or operational efficiency. It should also be judged by whether users can understand risk outputs, regulators can inspect evidence, organisations can assign responsibility, and human experts can challenge recommendations. These conditions are essential because food safety decisions affect public health, legal responsibility, and trust in food systems.

Embedding explainability into the next generation of food safety AI will require collaboration among AI developers, food businesses, regulators, auditors, and domain experts. Such collaboration can ensure that technical innovation is matched by transparent evidence, practical accountability, and meaningful human oversight. Explainable AI should therefore be treated as a foundation for trustworthy digital food safety governance rather than as an optional add-on.

## Acknowledgements

None

## Conflict of interest

None

## Financial support

None

## Ethics statement

None

Received: 23 Jan 2025

Revised: 06 Mar 2025

Accepted: 16 Apr 2025

Published online: 15 September 2025

## Rights and permissions

**Open Access** The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

## References

- Deng X, Cao S, Horn AL. Emerging applications of machine learning in food safety. *Annu Rev Food Sci Technol*. 2021;12:513-38. <https://doi.org/10.1146/annurev-food-071720-024112>.
- Marvin HJP, Janssen EM, Bouzembrak Y, Hendriksen PJM, Staats M. Big data in food safety: An overview. *Crit Rev Food Sci Nutr*. 2017;57(11):2286-95. <https://doi.org/10.1080/10408398.2016.1257481>.
- Bouzembrak Y, Klüche M, Gavai A, Marvin HJP. Internet of Things in food safety: Literature review and a bibliometric analysis. *Trends Food Sci Technol*. 2019;94:54-64. <https://doi.org/10.1016/j.tifs.2019.11.002>.
- Wang X, Bouzembrak Y, Oude Lansink AGJM, van der Fels-Klerx HJ. Application of machine learning to the monitoring and prediction of food safety: A review. *Compr Rev Food Sci Food Saf*. 2022;21(1):416-34. <https://doi.org/10.1111/1541-4337.12868>.
- Qian C, Murphy SI, Orsi RH, Wiedmann M. How can AI help improve food safety? *Annu Rev Food Sci Technol*. 2023;14:517-38. <https://doi.org/10.1146/annurev-food-060721-013815>.
- Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion*. 2020;58:82-115. <https://doi.org/10.1016/j.inffus.2019.12.012>.

Adadi A, Berrada M. Peeking inside the black-box: A survey on explainable artificial intelligence. *IEEE Access*. 2018;6:52138-60.  
<https://doi.org/10.1109/ACCESS.2018.2870052>.

Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206-15.  
<https://doi.org/10.1038/s42256-019-0048-x>.

Busuioc M. Accountable artificial intelligence: Holding algorithms to account. *Public Adm Rev*. 2021;81(5):825-36.  
<https://doi.org/10.1111/puar.13293>.

Martin K. Ethical implications and accountability of algorithms. *J Bus Ethics*. 2019;160(4):835-50.  
<https://doi.org/10.1007/s10551-018-3921-3>.

Mu W, Kleter GA, Bouzembrak Y, Dupouy E, Frewer LJ, Radwan Al Natour FN, et al. Making food systems more resilient to food safety risks by including artificial intelligence, big data, and Internet of Things into food safety early warning and emerging risk identification tools. *Compr Rev Food Sci Food Saf*. 2024;23(1):e13296.  
<https://doi.org/10.1111/1541-4337.13296>.

Rai A. Explainable AI: From black box to glass box. *J Acad Mark Sci*. 2020;48(1):137-41.  
<https://doi.org/10.1007/s11747-019-00710-5>.

Kaur D, Uslu S, Rittichier KJ, Durresi A. Trustworthy artificial intelligence: A review. *ACM Comput Surv*. 2022;55(2):39:1-38.  
<https://doi.org/10.1145/3491209>.

Nogales A, Díaz-Morón R, García-Tejedor ÁJ. A comparison of neural and non-neural machine learning models for food safety risk prediction with European Union RASFF data. *Food Control*. 2022;134:108697.  
<https://doi.org/10.1016/j.foodcont.2021.108697>.

Sari OF, Amer E, Bader-El-Den M, Ince V, Leadley C. AI-driven risk assessment in food safety using EU RASFF database. *Food Bioprocess Technol*. 2025;18(7):6282-303.  
<https://doi.org/10.1007/s11947-025-03819-4>.

Buyuktepe O, Catal C, Kar G, Bouzembrak Y, Marvin H, Gavai A. Food fraud detection using explainable artificial intelligence. *Expert Syst*. 2025;42(1):e13387.  
<https://doi.org/10.1111/exsy.13387>.

Magarelli M, Novielli P, De Filippis F, Magliulo R, Di Bitonto P, Diacono D, et al. Explainable artificial intelligence and microbiome data for food geographical origin: The Mozzarella di Bufala Campana PDO case of study. *Front Microbiol*. 2024;15:1393243.  
<https://doi.org/10.3389/fmicb.2024.1393243>.

Gunning D, Aha DW. DARPA's explainable artificial intelligence (XAI) program. *AI Mag*. 2019;40(2):44-58.  
<https://doi.org/10.1609/aimag.v40i2.2850>.

Sari OF, Bader-El-Den M, Leadley C, Esmeli R, Mohasseb A, Ince V. AI-driven food safety risk prediction: A transformer-based approach with RASFF database. *Br Food J*. 2025;127(9):3427-45.  
<https://doi.org/10.1108/BFJ-10-2024-1072>.

Röhrs S, Rohn S, Pfeifer Y. Risk classification of food incidents using a risk evaluation matrix for use in artificial intelligence-supported risk identification. *Foods*. 2024;13(22):3675.  
<https://doi.org/10.3390/foods13223675>.

Revelou PK, Tsakali E, Batrinou A, Strati IF. Applications of machine learning in food safety and HACCP monitoring of animal-source foods. *Foods*. 2025;14(6):922.  
<https://doi.org/10.3390/foods14060922>.

Marvin HJP, Bouzembrak Y, van der Fels-Klerx HJ, Kempenaar C, Veerkamp R, Chauhan A, et al. Digitalisation and artificial intelligence for sustainable food systems. *Trends Food Sci Technol*. 2022;120:344-8.  
<https://doi.org/10.1016/j.tifs.2022.01.020>.

Röhrs S, Nagy K, Kreutzer M, Stadler R, Rohn S, Pfeifer Y. Identifying food safety risks with a novel digitally-driven food safety early warning tool: A retrospective study on the pesticide ethylene oxide. *Food Control*. 2025;168:110939.  
<https://doi.org/10.1016/j.foodcont.2024.110939>.

Kroll JA. The fallacy of inscrutability. *Philos Trans A Math Phys Eng Sci*. 2018;376(2133):20180084.  
<https://doi.org/10.1098/rsta.2018.0084>.

Green B. The flaws of policies requiring human oversight of government algorithms. *Comput Law Secur Rev*. 2022;45:105681.  
<https://doi.org/10.1016/j.clsr.2022.105681>.

Lyons H, Velloso E, Miller T. Conceptualising contestability: Perspectives on contesting algorithmic decisions. *Proc ACM Hum-Comput Interact*. 2021;5(CSCW1):106:1-25.  
<https://doi.org/10.1145/3449180>.

Benefo EO, Karanth S, Pradhan AK. Applications of advanced data analytic techniques in food safety and risk assessment. *Curr Opin Food Sci*. 2022;48:100937.  
<https://doi.org/10.1016/j.cofs.2022.100937>.