GEORGE MASON UNIVERSITY

The Future of AI in Food Systems: Opportunities and Challenges for Human Technical Experts to Work Alongside AI Tools

Amarda Shehu

Professor of Computer Science, School of Computing

Associate Dean for Research, College of Engineering and Computing

Vice President and Chief AI Officer

JIFSAN-CFS3 Advisory Council Annual Symposium | College Park, Maryland · October 21, 2026

Central question for today: *under what conditions does adding an AI tool improve the work of a food-safety expert, and how would we know if it actually did?*

Case study: what 106 experiments tell us about people working with AI

Vaccaro, Almaatouq & Malone, Nature Human Behaviour, 2024.

Across 106 experiments, people, AI systems, and human–AI pairs performed tasks including detecting fakes, interpreting medical images, forecasting, classifying text, and writing.

How the results were measured

Hedges' g* measures the difference between two conditions relative to the variability in their scores. Values of 0.2, 0.5, and 0.8 were interpreted as small, medium, and large effects.

+0.64 Human–AI pair vs. human alone

On average, the pair performed better than the person working alone.

−0.23 Human–AI pair vs. the better individual

The pair performed worse than whichever member, human/AI, performed better alone.

−0.27 Decision tasks

Largest difficulty appeared in tasks that required choosing from predefined options.

Putting a human in the loop on a well-specified classification task cost accuracy relative to the model by itself, because the human could not tell which of the model's answers to trust.

Instructive for food safety. *Triage, risk ranking, and evidence curation are decision tasks. Improving on unaided human performance does not establish that a team is better than its strongest member.*

Important distinction to evaluate an AI-assisted workflow.

*Hedges' g is a standardized effect size. It expresses the difference between two group means in units of their pooled standard deviation, allowing comparisons of studies with different scales and different outcome measures.

Does AI strengthen an expert team? Four guiding criteria/questions.

1. Independence

Can the expert catch the tool's mistakes?

2. Calibration

Can we tell when the tool is reliable?

3. Observability

Can we measure what a deployed system misses?

4. Loss

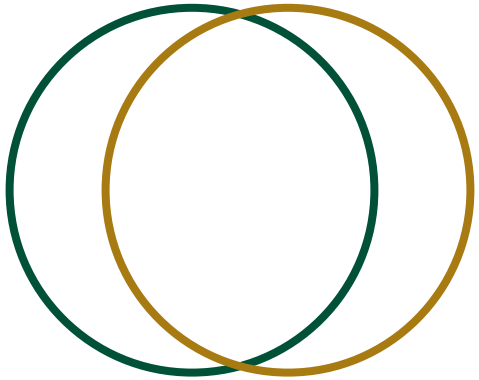
Can we make the consequences of a decision explicit?

These questions depend on one another:

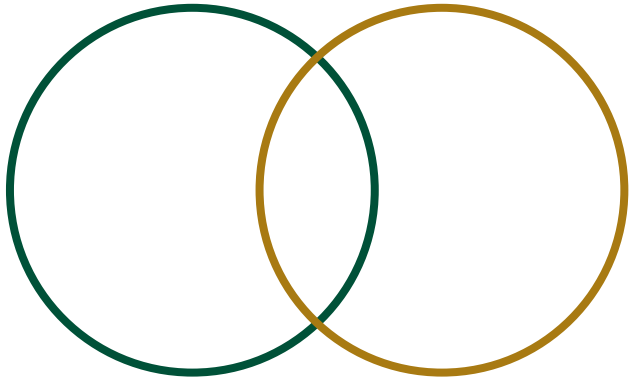
- Independence concerns which errors a team can correct.
- Calibration helps determine when to rely on the tool (and when not).
- Observability determines whether we can measure performance after deployment.
- Loss determines which errors count most when a decision is made.

1. Independence: Can the expert catch the tool's mistakes?

Green: expert errors



Gold: tool errors



Shared errors

When expert and tool fail on the same cases, combining them offers little protection.

Distinct errors

When their mistakes differ, one member can correct the other.

Ensemble error

For an ensemble, error can be expressed as average member error minus a diversity term. Diversity reflects disagreement among members. *Krogh & Vedelsby, 1995.*

Practical evaluation question

On a common set of cases, how often is the expert wrong, how often is the tool wrong, and how often are they both wrong?

> *Accuracy tells us how often a tool is right. Error overlap tells us whether it can contribute something (complementary that) the expert does not already know.*

Warning: A model can reproduce an expert's blind spots

Training target	What the model learns	Where its errors may overlap with expert errors
Past analyst decisions	How analysts made earlier decisions	By construction, it learns to reproduce those decisions
Inspector-recorded violations	Patterns in recorded inspection outcomes	Through the inspector's observations *and recording practices*
Third-party audit scores	Patterns in audit assessments	Through the checklist *and its omissions*
Laboratory results linked to whole-genome sequencing clusters	Evidence closer to underlying hazard	Potentially greater independence from inspection and audit judgments

Important to ask: *What exactly was the model trained to predict, and how was that target established?*

A model trained to reproduce expert decisions may apply them consistently or at scale, but it cannot be assumed to identify hazards the experts overlook(ed) in the first place.

Case study: Chicago's restaurant inspection model

The Chicago Department of Public Health developed a model to rank approximately 15,000 establishments by their predicted likelihood of a critical food-code violation at the next routine inspection. Inspectors visited higher-scoring establishments first.

Researchers replayed two months of actual inspections in the model's proposed order and compared the results with the city's actual inspection order. A critical violation is one recorded as serious enough to require correction. *Kannan et al., 2019.*

7.5 days earlier detection

On average, a critical violation was identified sooner under the model's proposed ordering.

69% vs. 55% violators found in first half

Model ordering vs. actual inspection order.

Odds ratio 4.7 Inspector-group variation

One inspector group had 4.7 times the predicted odds of recording a critical violation, regardless of the establishment.

What this case teaches us. *The model identified a useful signal, but its training labels were inspectors' records, which reflected both conditions at establishments and the people who inspected them.*

The retrospective gains do not, by themselves, tell us how much model errors overlapped with human errors.

Warning: AI systems that check each-other can have correlated errors

Researchers compared hundreds of language models answering the same multiple-choice questions.

Evaluation included 349 models on 12,000 questions and 71 models on 14,000 questions.

They also used 20 models to screen 1,800 résumés. *Kim, Garg, Peng & Garg, ICML 2025.*

About 60% shared wrong answers

When two models were wrong, they gave the same wrong answer approximately 60% of the time. With 4 answer options and unrelated errors, expected rate would be ~25%.

Higher accuracy, greater error agreement

More accurate models also tended to agree more often when they were wrong, including models from different companies.

Grader bias

A model evaluating other models tended to favor those that shared its own incorrect answers. [Not independent check]

Implication for food-safety workflows. *A second model checking an extraction, classification, or summary is not necessarily an independent check.*

For scientific evidence, the underlying manuscript, laboratory record, or a qualified human review provides a different, complementary, independent source of verification.

Lesson #1: Keep human judgment measurable after AI enters workflow

A simple evaluation protocol

Periodically, ask experts to assess a sample of cases without seeing the tool's output (no anchoring).

For inspection work, sanitarians can independently rescore a sample of reports.

For evidence curation, analysts can classify a sample of complaints or findings without AI assistance.

What to measure

The expert's *unaided* accuracy.

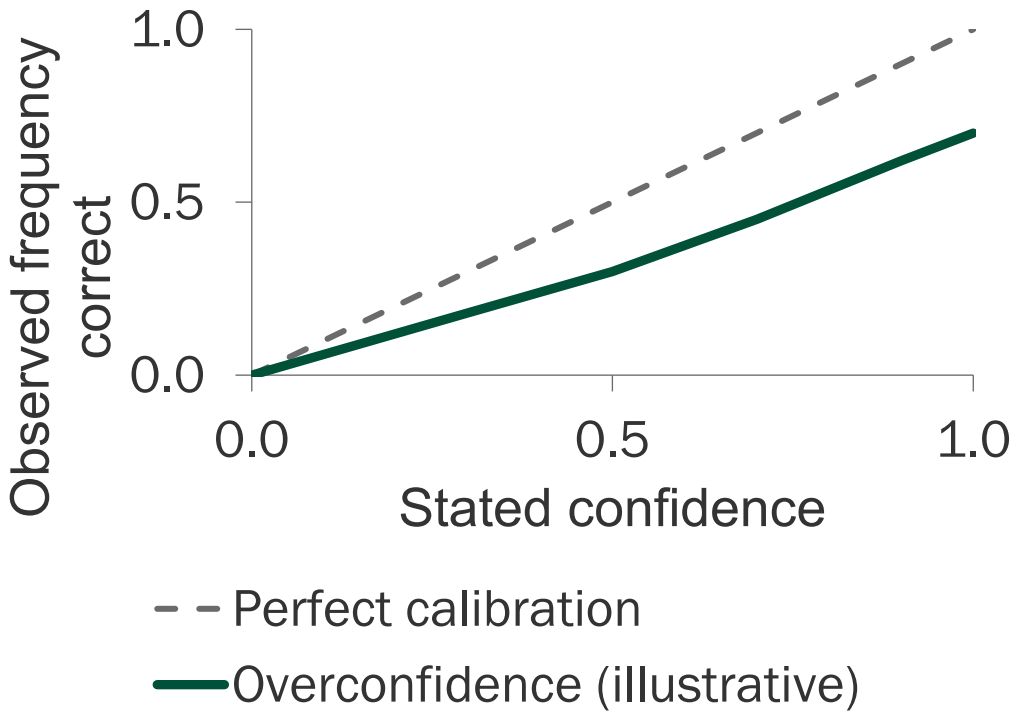
Overlap between expert and tool errors.

Whether the relationship changes as staff gain experience with the tool.

If human and tool errors become more alike over time, the team may be losing its ability to catch mistakes even while the dashboard's performance figures remain unchanged.

Operational question. *Does the workflow preserve independent human judgment, and (how) is that independence being (regularly) measured?*

2. Calibration: When a tool says 90%, is it right about 9/10 times?



Reading the chart

Perfect calibration: observed frequency matches the stated confidence.

Overconfidence: observed frequency falls below it. Among cases assigned 90% confidence, a calibrated system is correct about 90% of the time across a large and relevant set of cases.

Why calibration follows independence

Independence tells us whether a human *could* correct the tool. Calibration helps us decide *when*. Both require observed outcomes over confidence.

A confidence score becomes useful for decision-making only when we have evidence about what it means on relevant cases

Case study: conversational fluency is not evidence of calibrated confidence

Study #1: Confidence before and after conversational training

OpenAI, 2023. For one model answering exam questions, calibration error increased from 0.007 before post-training to 0.074 afterward, a roughly 10x increase.

Calibration error measures the difference between stated confidence and observed accuracy; zero indicates perfect calibration under used measure.

Study #2: How people use AI assistance

In an experiment involving 1,626 participants, people assessed reviews with AI assistance. Showing confidence information produced team accuracy of 0.89; adding explanations produced 0.88, with no improvement.

Explanations increased reliance on the AI, including when its answers were wrong. *Bansal et al., CHI 2021.*

Gullibility for food safety. *Nutrition information, evidence synthesis, and scientific curation increasingly use conversational AI. A persuasive explanation is not a reliability test. Confidence and accuracy must be evaluated on the kinds of cases the system will encounter in practice.*

From my own undergraduate students in AI4All (UNIV 182)
The midterm: three weeks, ten prompts, four chatbots

1 Design

Author 10 prompts across 6 reasoning families

2 Compare

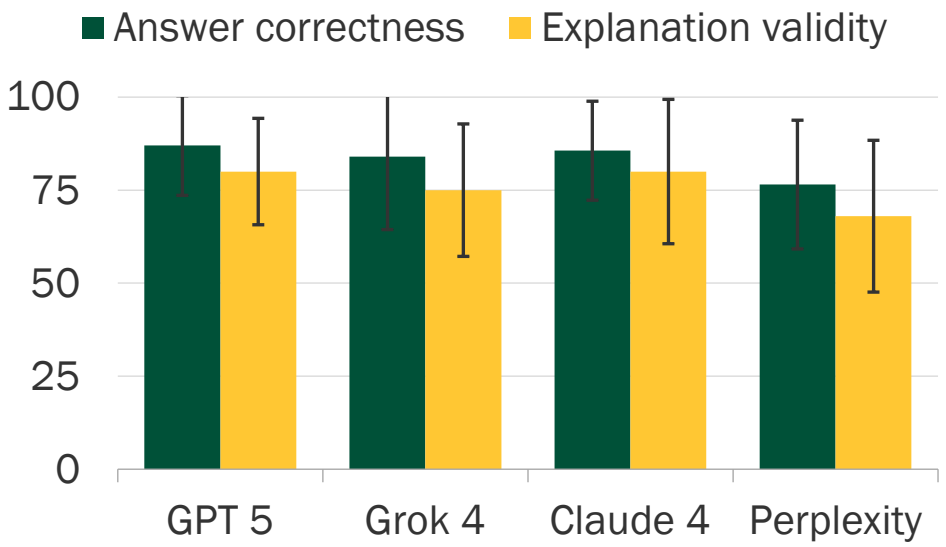
Submit identical wording to 4 chatbots

3 Score

Rate the answer and the explanation separately

4 Reconcile

Resolve scoring disagreements and report evidence



For each, mean explanation validity fell below mean answer correctness, by 5.6–9.0 points.

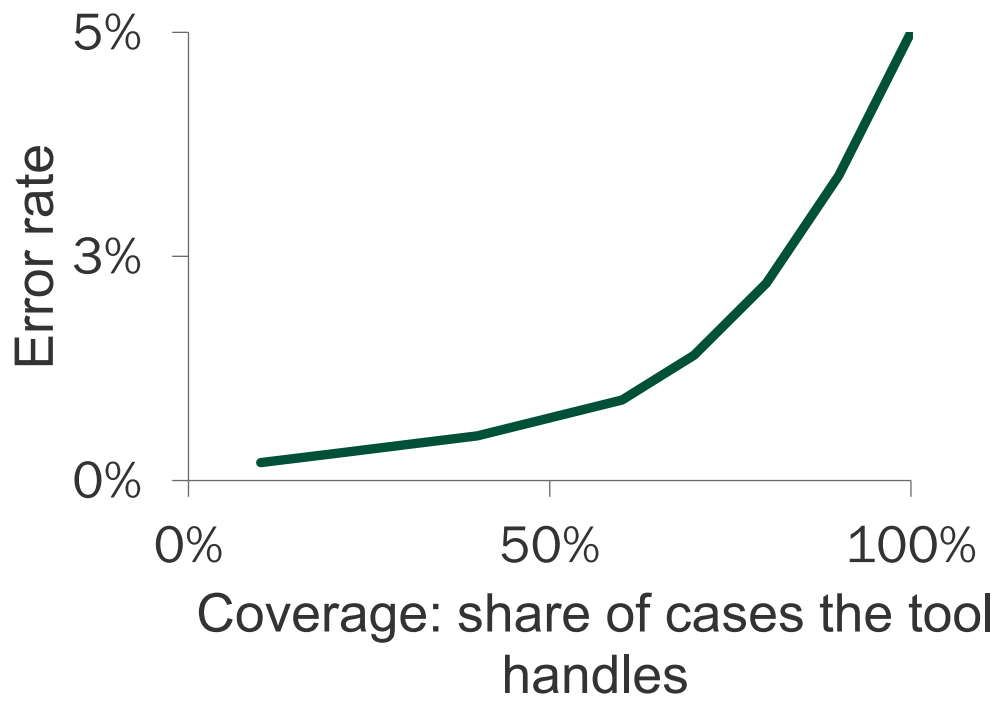
"Good pattern followers, not reliable general thinkers."

"Users cannot trust confidence or nice wording as proof the reasoning is correct."

"If we had not checked, we would have trusted it." Student teams, UNIV 182

Shehu et al., "AI Literacy as Experimental Practice: Students as Investigators," *Communications of the ACM 69 (9): 48-56,* 2026 (with all students as co-authors)

Lesson #2: Measure error at each level of coverage



An example operating agreement

Tool handles 60% of cases for which its predictions meet a prespecified error requirement, below 1%. Experts review remaining 40%, where tool is at most advisory. (percentages are illustrative; a threshold must be set on relevant evaluation data).

Conformal prediction

Statistical framework that wraps a prediction model to provide specified coverage guarantees, subject to assumptions about how evaluation cases relate to future cases.

The operational point. *A single accuracy score averages over cases the tool handles well and cases where it may be unreliable. Risk–coverage analysis exposes this distinction.*

Changes in the food system can invalidate a model's reliability

Type of change	What changes	Food-safety example	Who may recognize it
Covariate shift	The distribution of inputs	New suppliers, products, or import regions	Sourcing and supply-chain experts
Label shift	The prevalence of outcomes	A seasonal rise in temperature abuse, or an outbreak that raises hazard prevalence	Epidemiologists and surveillance teams
Concept shift	The relationship between inputs and outcomes	A pathogen–product pairing not represented in the model's history	Microbiologists with knowledge of the mechanism

Operational requirement. *A system may detect that its inputs look unfamiliar without recognizing that a familiar input now carries a different risk. Determine who is responsible for recognizing each type of change, what evidence triggers reassessment, and when calibration must be checked again.*

Case study: Listeria in whole cantaloupe

Jensen Farms

In 2011, whole cantaloupes from a Colorado farm were linked to an outbreak across multiple states. Investigators found outbreak strains on packing equipment used for another crop and documented pooled water and the absence of pre-cooling. *CDC, 2012; McCollum et al., NEJM, 2013.*

Before outbreak, whole cantaloupe was not a familiar source of Listeria monocytogenes. A model trained on historical product–pathogen associations might assign the pairing a low risk even when physical conditions support contamination and growth.

147 illnesses 33 deaths

Outbreak involved 28 states. The median age of those who fell ill was 78. It was among the deadliest foodborne outbreaks in decades.

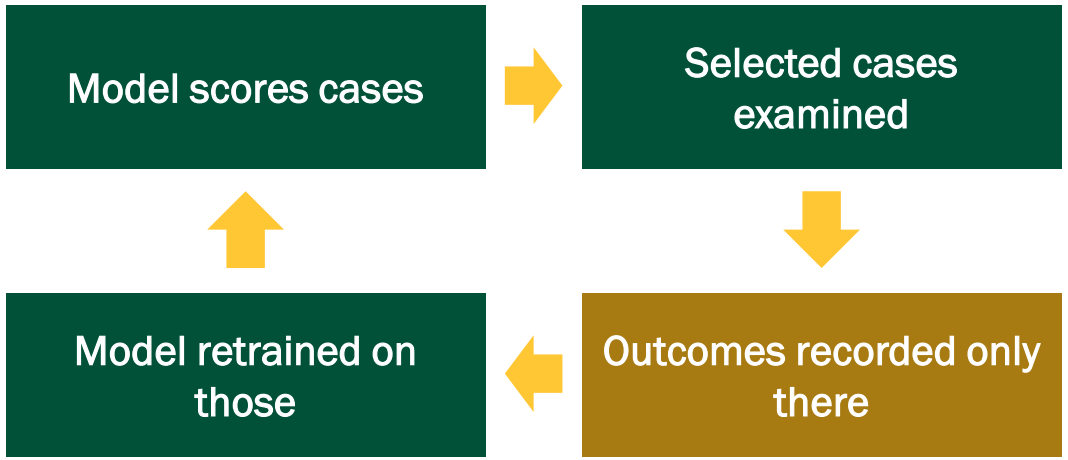
Scientific lesson

A netted rind can retain moisture, and contaminated wet environments can support microbial growth. Understanding those mechanisms provides a way to recognize hazards that a model's historical examples cannot capture.

The facility had passed an audit days before outbreak.

Independent knowledge of the hazard remains essential.

3. Observability: Can we measure what deployed system misses?



Selective labels. Outcomes are recorded only for examined cases, so the system's own data cannot establish how often it misses hazards among cases it never selected.

Performative prediction

Predictions influence the outcomes being predicted. Facilities targeted for inspection may change their practices, while inspectors may examine them more closely. The data reflect both the conditions and the intervention.

Evaluation problem

If a model is retrained on outcomes collected mainly from cases it selected, its future training data inherit selection pattern of earlier predictions.

After deployment, a system changes which evidence becomes available.

To evaluate a deployed model, we need some source of outcome data beyond the cases that the model itself directs us to examine.

Food safety already faces the selective-observation problem

Import screening

A score determines which shipments are examined. Findings from those inspections do not, by themselves, reveal the hazard rate among shipments that were not examined.

Inspection scheduling

In Chicago, violations were observed where inspections occurred, and the records reflected inspectors' observations. The model then helped determine which establishments would be inspected.

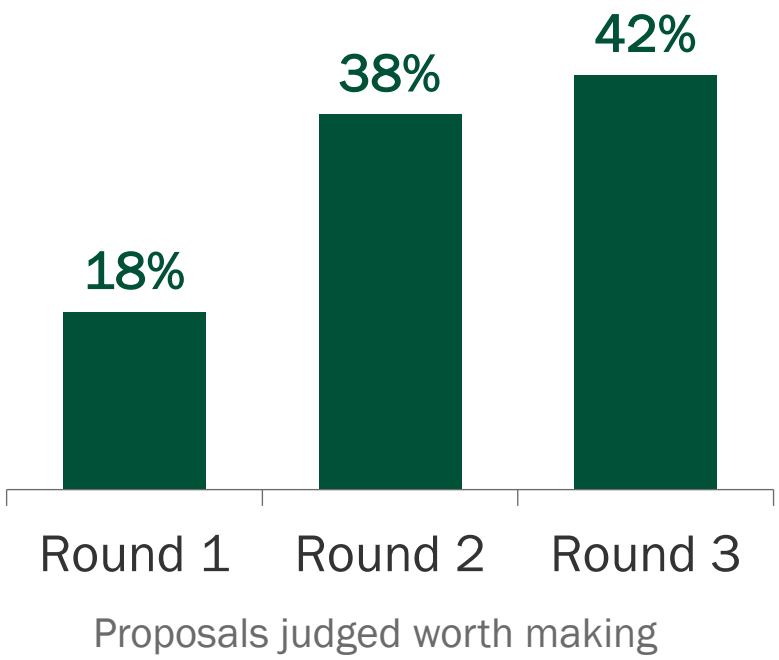
Complaint triage

A score determines which complaints receive investigation. The uninvestigated complaints lack equivalent outcome data.

Without an independent way to select cases for examination, a model determines the outcomes and so the evidence on which it may/will be evaluated.

Practical safeguard. *Food-safety programs already use baseline and surveillance sampling that is independent of risk scores. Preserve a defined portion of that sampling when introducing AI.*

Lesson #3: When evidence comes from outside the model, the feedback loop becomes useful



Example: AI-assisted discovery of sanitizing compounds

Quaternary ammonium compounds (QACs) are active ingredients in many sanitizers. The model proposed molecular structures, and chemists assessed each proposal: worth making, already existed, or not a QAC.

A chemical reality check

Across rounds 1–3, the share of proposals that already existed fell from 47% to approximately zero. One candidate inhibited Staphylococcus aureus at 1 μM, compared with 2 μM for benzalkonium chloride, a commonly used QAC. *Shehu lab, KDD 2025.*

Why external judgment is crucial. *The model could generate and rank candidates, but chemical identity, novelty, and biological activity required independent evidence beyond its own predictions. Chemists supplied the independent assessments that allowed the model to improve upon rounds.*

Budget for exploration so that missed hazards remain visible

Estimating the required sample

A common approximate sample size for a proportion is

$$n = z^2\, p(1 - p) / E^2$$

where p is the expected miss rate, E the desired margin of error, and z = 1.96 for 95% confidence.

For $p \approx 2\%$ and E = one percentage point, this gives about 753 randomly selected examinations per estimate. The calculation is an approximation; the required sample also depends on the sampling design.

What independent sampling provides

An estimate of the miss rate among cases the model would not ordinarily select (below the threshold).

Evidence for checking calibration against observed outcomes.

An early indication that conditions may have changed.

Operational recommendation. *Reserve a defined budget for randomly selected examinations, including cases below the model's action threshold. This is how a food-safety program can learn about hazards its predictive system may be overlooking.*

4. Loss: Every action threshold encodes a judgment about risk

Action threshold	Implied cost of a miss relative to a false alarm
0.50	1×
0.20	4×
0.10	9×
0.05	19×
0.01	99×

These ratios follow from the simplified rule at right. They are not estimates of actual food-safety costs.

When should the system recommend action?

Act when the estimated probability of a hazard exceeds the threshold implied by the relative costs of a false alarm and a missed hazard:

$$t = C_{FP} / (C_{FP} + C_{FN})$$

C_FP is the cost of a false positive; C_FN is the cost of a false negative.

Many systems are tuned to maximize a validation metric. That choice does not establish whether threshold is appropriate for consequences of a decision.

Question for deployment. *Who chose the threshold, what relative costs does it imply, and who is accountable for that choice?*

Food-safety decisions have different costs (even when they use the same score)

Recall

A missed Listeria signal in a ready-to-eat product may expose consumers to serious harm.

Public advisory

A missed undeclared allergen may leave consumers without information needed to avoid unsafe product.

Inspection

A misdirected visit consumes inspection time that could have been spent elsewhere.

Hold

Holding a lot without sufficient cause can delay distribution and result in lost product.

The same model score may inform all four actions, but consequences of false alarms and missed hazards differ.

Requirement. *Each action needs an appropriate threshold, an explicit account of the relative costs, and a designated person responsible for approving the decision rule. Recording those choices makes the basis of an AI-assisted decision (actually) open to (human) review.*

Lesson #4: Evaluate human–AI team at threshold used in practice

Expected cost: Estimate cost of errors at deployed decision threshold, using stated consequences of false positives and false negatives.

Team versus best member: Compare human–AI team with the better of the expert working alone and the tool working alone, on same cases and evaluation criteria.

Measured performance versus perceived benefit: Ask staff whether they believe the tool helps but do not treat perceptions as substitute for measured performance. Controlled studies report cases where perceived and measured benefits had opposite signs.

Why not rely on AUC alone?

The area under the receiver operating characteristic curve summarizes discrimination across thresholds.

It does not tell us the cost of errors at the threshold used for a particular food-safety action.

A tool should be evaluated in the workflow where it will operate. If the combined arrangement performs worse than its strongest individual component under the relevant loss measure, the arrangement needs to be reconsidered.

A one-page specification for evaluating an AI tool in food safety

Item	Question to ask	Quantity
Training target	What is the model predicting, and whose judgment appears in its labels?	Independence
Error overlap	How often are the tool and the expert wrong on the same cases?	Independence
Calibration	How reliable are its confidence estimates on current data and what triggers recalibration?	Calibration
Risk–coverage	What is the error rate at each level of coverage, and where will the tool operate?	Calibration
Exploration	What proportion of cases is examined independently, and what miss rate does that sample reveal?	Observability
Threshold and loss	What relative costs does each action threshold imply, and who approved it?	Loss

Lessons in Practice: Chicago revisited

Independence

Model learned from inspectors' records. How often did its errors coincide with the inspectors' errors?

Calibration

Did the relationship between predicted risk and observed violations hold across seasons and changing conditions? Was that checked?

Observability

Model learned from establishments inspectors had chosen to visit, then helped choose which would be visited. How were outcomes measured among those not selected?

Loss

What was the cost of missing a critical violation, and how was that cost reflected in the threshold for inspection?

Working alongside AI requires evidence about the team

Thank you