

ORIGINAL RESEARCH

Open access

Managing Hallucination Risk in Agricultural AI: A Safety Framework for Generative Decision Support in Precision Farming

Daniel Kim^{1*}, Jihoon Park¹, Min-seo Kang², Hyun-woo Jung³

Abstract

Generative artificial intelligence is expanding the conversational and interpretive functions available within precision-farming decision support, yet fluent responses may contain factual inaccuracies, unsupported claims, fabricated evidence, agronomically invalid interpretations, local-context mismatches, or understated uncertainty. This article develops a conceptual artificial-intelligence safety framework for managing these risks in generative agricultural decision-support systems. The framework treats hallucination risk as a socio-technical safety problem rather than solely as a language-model defect. It links agricultural use-case definition, advisory-output classification, farm-level consequence assessment, data and knowledge provenance, retrieval and response grounding, hallucination screening, factual verification, agronomic-validity review, uncertainty communication, human oversight, expert escalation, farmer feedback, traceability, accountability, monitoring, and controlled updating. Generative outputs are differentiated according to whether they provide information, interpretation, scenario explanation, or recommendation-like content, because the evidentiary and review requirements increase with potential consequence. Validation is similarly separated into factual checking, external evaluation, local-context assessment, agronomic review, and future field-validation planning. The framework establishes decision boundaries against unsupported agronomic advice, autonomous farm action, hallucination-elimination claims, safety guarantees, and deployment-readiness claims. It also allocates responsibilities among developers, platform providers, data stewards, knowledge-base curators, agricultural reviewers, and farmers while preserving farmer autonomy. The main contribution is an integrated safety model for structuring how hallucination-prone generative outputs may be interpreted, reviewed, restricted, escalated, recorded, and updated within precision-farming decision support. The framework is conceptual and does not independently establish advisory correctness, operational safety, legal compliance, or deployment readiness.

Keywords Precision farming, Decision support, Generative AI, Agricultural AI, Hallucination, AI safety

*Correspondence:

Daniel Kim

daniel.kim@gmail.com

¹ Department of AI Safety and Hallucination Management, College of Agriculture, Seoul National University, Seoul, South Korea

² Department of Generative Decision Support, College of Agriculture, Chung-Ang University, Seoul, South Korea

³ Department of Agricultural AI Risk Assessment, College of Agriculture, Kangwon National University, Chuncheon, South Korea

Introduction

Precision farming increasingly combines sensor observations, weather information, soil and crop data, remote-sensing products, farm-management records, analytical models, and digital platforms to support agricultural decisions. These systems are not purely technical artefacts: they operate within agricultural knowledge networks, institutional arrangements, commercial relationships, farmer practices, and locally specific production environments. Their outputs

therefore depend not only on model capability but also on data quality, task definition, system integration, stakeholder roles, and the conditions under which information is interpreted and used [1-3]. The introduction of large language models, foundation models, vision-language systems, and retrieval-supported generation adds a conversational layer capable of synthesising information and explaining apparent implications, but it also creates new routes through which unsupported or misleading content can enter farm decision-support workflows.

Hallucination risk is especially significant when generated language resembles authoritative farm advice. A response may be grammatically coherent and contextually persuasive while containing an incorrect factual statement, an unsupported inference, fabricated evidence, an invalid agronomic interpretation, or an assumption that does not fit the farm's crop, soil, climate, management history, or production objective. These failure forms should not be collapsed into one undifferentiated category. Hallucination concerns content generated without adequate grounding; factual inaccuracy concerns claims that conflict with reliable evidence; agronomic invalidity concerns interpretations or recommendations that are biologically or operationally unsuitable; and local-context mismatch concerns outputs that may be broadly plausible but inappropriate for the intended setting. Model reliability refers to the consistency of system behaviour, whereas model validity concerns whether an evaluated system is suitable for a specified task, population, environment, and decision purpose.

Agricultural AI safety consequently requires more than model-performance reporting. Responsible use depends on reliability assessment, transparent data practices, documentation of limitations, meaningful human involvement, accountability, privacy protection, cybersecurity, and safeguards against harmful or misleading recommendations [4]. These requirements become more demanding when outputs influence decisions concerning irrigation, nutrients, pests and diseases, crop monitoring, weather-related risk, or other resource-sensitive activities. A general-purpose model may assist with information organisation or explanation, but its apparent confidence does not establish that its output is current, locally valid, supported by traceable evidence, or suitable for consequential use. Agricultural decision support must therefore distinguish between information provision and advice-like outputs whose interpretation could affect farm resources, production conditions, or user behaviour.

This article proposes an original conceptual AI safety framework for managing hallucination risk in generative agricultural decision support. The framework links hallucination-risk identification, advisory-output classification, farm-level consequence analysis, evidence grounding, factual and agronomic review, uncertainty communication, human oversight, expert escalation, farmer confirmation, decision boundaries, accountability allocation, traceability, monitoring, incident learning, and controlled updating. Its purpose is not to establish a universal validation protocol or prescribe farm-management actions. Instead, it structures the questions, responsibilities, review gates, and claim limitations that should be considered before a generated output is treated as decision support. The central argument is that generative agricultural systems should be judged by the quality of their evidence, validation, oversight, and boundary enforcement rather than by conversational fluency, response speed, or general model capability [4].

Hallucination Risk in Agricultural AI

Hallucination in generative AI refers broadly to content that is unsupported by the relevant source material, supplied data, or established evidence, although its manifestation varies across dialogue, question answering, summarisation, data-to-text generation, and multimodal systems. In an agricultural setting, hallucination may appear as an invented observation, an unsupported causal explanation, a fabricated source, a claim that exceeds retrieved evidence, or an answer that fills missing farm context with unstated assumptions. Generated citations can appear credible while referring to non-existent or misrepresented sources, and polished conversational language can create an impression of truth-sensitive reasoning that the generation process does not necessarily possess [5-7]. Hallucination should therefore be separated from ordinary imprecision, incomplete explanation, model unreliability, and invalid agronomic judgement, even though these failure categories can occur together in the same output.

Factual inaccuracy concerns whether a generated statement is correct, while unsupported output concerns whether the statement can be traced to accepted evidence. A claim may be factually correct by coincidence yet remain unsupported within the actual decision-support workflow. Conversely, an output may cite a genuine source but misstate its findings or extend them beyond the context in which they are valid. Fabricated evidence and fabricated source attribution are especially serious because they give unsupported content an appearance of verification [6]. Overconfident output adds a second layer of risk by presenting weak, incomplete, or conflicting evidence in categorical language. Because users often rely on linguistic cues when judging credibility, confident wording can obscure uncertainty, discourage further checking, and increase the likelihood that informational content will be interpreted as actionable advice [7].

Agricultural hallucination risk is also produced upstream of text generation. Poor sensor calibration, incomplete farm records, incorrect field boundaries, stale weather data, low-quality imagery, uncertain soil measurements, or missing provenance can distort the context supplied to the model. Ambiguous farmer queries may omit crop variety, location, growth stage, production objective, timeframe, prior intervention, or other material conditions, encouraging the model to generate assumptions rather than request clarification. Knowledge-base drift may occur when curated agricultural sources become outdated, inconsistent, or incorrectly indexed, while model drift may arise after model, prompt, retrieval, or interface changes. Agronomic invalidity and local-context mismatch may remain even where retrieved facts are individually correct, because agricultural decisions depend on interactions among biological conditions, management constraints, farmer experience, local knowledge, and temporal circumstances [5].

Retrieval-augmented generation can help connect outputs to external agricultural knowledge, but it does not remove hallucination risk. Retrieval may return irrelevant, outdated, conflicting, or geographically inappropriate evidence; ranking may prioritise superficially similar material; and the generation stage may ignore, distort, or overextend the retrieved context. Retrieval quality and response grounding must therefore be evaluated separately, with claim-level links between generated statements and supporting evidence. Hallucination detection, source checking, knowledge-base curation, and human review may reduce selected risks where they are properly designed and validated, but none creates a hallucination-free system [8]. An explicit no-hallucination-elimination boundary and no-advisory-correctness-guarantee boundary are thus necessary: passing proposed checks may support more cautious interpretation, but it cannot independently guarantee factual completeness, agronomic validity, local suitability, operational safety, or deployment readiness.

Generative Decision Support and Farm-Level Consequences

Generative decision support can translate complex agricultural information into conversational explanations, summaries, comparisons, clarification questions, and scenario descriptions. In crop production, large language models may provide value when they help organise textual knowledge, make technical information more accessible, or support interaction with broader analytical systems, but their appropriate role depends on the task, the evidence available, and the consequence of error [9]. A safety framework should therefore classify outputs before evaluating them. Information retrieval, explanation, interpretation, scenario comparison, and recommendation-like language should not be treated as equivalent. As an output moves closer to a proposed farm-management action, stronger grounding, validation, human review, and decision-boundary controls become necessary.

Generative systems may be connected conceptually to sensor data, weather services, remote-sensing products, farm records, and analytical models, allowing them to explain patterns across heterogeneous inputs. Such integration can improve accessibility but also compounds failure pathways because incorrect data, weak retrieval, model error, and misleading language can be combined into a single persuasive response [10]. For irrigation, nutrient, pest, disease, weather-risk, crop-monitoring, or remote-sensing questions, the system should not infer that access to one data stream is sufficient for advice. Soil moisture alone, for example, does not represent the full decision context for irrigation; an image-

based pattern does not by itself establish a diagnosis; and a forecast does not guarantee a particular farm outcome. These are decision boundaries, not agronomic instructions.

Agricultural knowledge bases and retrieval systems may support more grounded generative outputs by providing domain-specific evidence, terminology, and contextual information. However, useful grounding requires curated sources, transparent inclusion criteria, version control, relevance assessment, and evaluation of whether the retrieved material fits the crop, region, season, and intended decision [11]. Farmer-facing explanations should disclose material assumptions, missing data, uncertainty, and alternative interpretations in understandable language. Advisor-facing outputs may require more detailed evidence maps showing data provenance, source alignment, model version, unresolved conflicts, and the reasons an output was classified at a particular risk level. Neither form should obscure uncertainty merely to improve readability.

Farm-level consequences depend on how an output is interpreted and whether it crosses from explanation into prescription. A misleading scenario explanation may distort risk perception; an unsupported interpretation of remote-sensing data may direct attention away from relevant field observations; and recommendation-like wording may be acted upon even when the system labels itself as advisory. Integrating generative models with crop-management decision systems therefore requires staged evaluation, local assessment, and strict separation between conversational generation and autonomous control [12]. High-risk or insufficiently grounded queries should be deferred to qualified review. Human oversight must occur before consequential reliance, and no generated output should directly trigger machinery, resource application, or another farm-management action solely because it has passed automated screening.

Risk Interpretation, Validation, and Human Review

Risk interpretation begins by classifying the intended output according to potential consequence, reversibility, uncertainty, and required expertise. A low-consequence informational explanation does not require the same review as recommendation-like content that could influence resource use or crop management. Human-centred AI design shows that high automation need not require low human control; a system may assist with information processing while preserving meaningful human authority, contestability, and responsibility [13]. Within agricultural decision support, this means that farmers and qualified reviewers must be able to inspect assumptions, challenge generated claims, reject outputs, request more evidence, and prevent automated execution. Human involvement must be substantive rather than ceremonial, because a nominal review step can otherwise become a mechanism for transferring responsibility without improving safety.

Validation should be separated into complementary functions. Factual accuracy checking asks whether claims agree with reliable and current evidence. Agronomic-validity review asks whether the interpretation is biologically, environmentally, and operationally coherent. External validation examines performance outside development data, while local validation considers the intended crops, soils, climate, language, users, and workflows. Field-validation planning specifies what future empirical evaluation would require but does not constitute completed field validation. Explainability can support these processes by exposing influential variables, model reasoning structures, or relationships among data and outputs, yet an explanation is not proof that the underlying claim is correct or agronomically suitable [14]. An understandable error remains an error, and a plausible explanation can itself be misleading.

Local and experiential knowledge are essential because agricultural knowledge cannot be reduced completely to digitised records or general textual sources. Farmers, agronomists, extension professionals, and other agricultural knowledge actors may recognise seasonal, spatial, historical, or operational conditions that are absent from the system's data [15]. Farmer feedback may identify an incorrect assumption, a missing constraint, or a mismatch between the output and observed field conditions. Agricultural advisor review may help evaluate consequence and contextual validity where relevant, but neither farmer confirmation nor expert participation guarantees correctness. Review quality depends on competence, authority, time, access to traceable evidence, awareness of automation bias, and the ability to defer or reject the output.

Validation and review should therefore be embedded within an iterative responsible-development process involving anticipation, stakeholder inclusion, reflection, responsiveness, monitoring, and revision [16]. Consequential outputs require traceable records of inputs, retrieved evidence, model version, generated claims, uncertainty disclosures, escalation decisions, and human review. Monitoring should examine model and knowledge-base drift, recurrent grounding failures, inappropriate confidence, boundary violations, and disagreement between automated and human assessments. A no-validation-sufficiency boundary is necessary because successful evaluation in one setting does not demonstrate universal reliability, and a no-deployment-readiness boundary is necessary because a conceptual safety framework does not establish operational capacity, field validity, organisational assurance, or safe real-world performance. **Table 1** summarises hallucination risks, generative decision-support consequences, validation needs, and human-review requirements for agricultural AI.

Table 1. Hallucination Risk, Generative Decision Support, Farm-Level Consequences, Validation, and Human Review in Agricultural AI: Factual Inaccuracy, Unsupported Outputs, Fabricated Evidence, Agronomic Invalidity, Local-Context Mismatch, Uncertainty Underreporting, Risk Classification, Human Oversight, Expert Escalation, Farmer Feedback, and Decision Boundaries

Risk, consequence, or review component	Core safety question	Potential source or required evidence	Risk manifestation or safety function	Potential farm-level consequence or output	Validation or review need	Residual failure
Factual inaccuracy	Is each generated factual claim correct and current?	Farm inputs, traceable agricultural sources, timestamps, and contextual metadata	Incorrect statement about a crop, condition, observation, or evidence source	Misinterpretation of the farm situation or evidence base	Independent factual verification and source checking	Correct f still be in or agron irrele
Unsupported output	Can each material claim be linked to accepted data or evidence?	Claim-source map, retrieved passages, supplied records, and provenance	Plausible statement generated without adequate support	Speculation may be interpreted as established knowledge	Claim-level grounding review	An outp appear c while exc sou
Fabricated evidence	Has the system invented a measurement, observation, document, or finding?	Original records, source identifiers, and retrieval logs	Non-existent evidence presented as real	False confidence in an interpretation or decision	Evidence existence and authenticity checks	Subtle fa may e autor dete
Fabricated source attribution	Do cited sources exist and support the attributed claim?	Verified article metadata, DOI, authorship, and relevant text	Invented or misrepresented citation	Unsupported advice may appear authoritative	Bibliographic and claim-level verification	Genuine may still l inaccu
Overconfident output	Is certainty stronger than the underlying	Evidence quality, model limits, conflicting sources, and missing data	Definitive language under uncertainty	Reduced likelihood of further checking or escalation	Confidence-language and evidence-strength review	Users r over-trus wor

Data provenance gap	Can each input and transformation be reconstructed?	Source, ownership, timestamp, processing, and version metadata	Untraceable data used in a consequential output	Failure cannot be investigated or responsibility allocated	Provenance audit	Prove record documents correct workflow
Ambiguous farmer query	Is sufficient context available to answer the question?	Crop, location, growth stage, objective, timeframe, and prior actions	Model fills omitted context with assumptions	Apparently specific but contextually unsupported response	Clarification request and assumption review	User-provided context may be incomplete
Model drift	Has behaviour changed following model, prompt, tool, or interface updates?	Version records, regression tests, and monitored output characteristics	Previously evaluated controls no longer behave consistently	New or recurrent failure patterns	Version-specific re-evaluation	Testing coverage conditions
Knowledge-base drift	Has the grounding corpus become stale, inconsistent, or corrupted?	Curation records, source conflicts, indexing, and update history	Retrieval quality deteriorates despite unchanged model	Outdated or contradictory decision support	Curator review and version control	Curated knowledge may not reflect current information
Risk classification	How consequential, reversible, and uncertain is potential reliance?	Intended use, decision type, affected resources, severity, and reversibility	Assigns control intensity and review requirements	Informational release, restricted output, escalation, or prohibition	Multidisciplinary classification review	Underclassification may be unsafe
Consequence severity assessment	What could happen if the output is wrong or misunderstood?	Crop, environmental, economic, operational, and human implications	Evaluates scope, duration, and reversibility	Consequence profile	Domain and stakeholder review	Indirect consequences may be overlooked
External validation	Does performance extend beyond development conditions?	Independent data and predefined evaluation criteria	Tests generalisation	Bounded validation statement	Methodological review	External conditions may not match intended conditions
Local validation	Is the system suitable for the intended agricultural setting?	Local crops, soils, climate, language, users, and workflows	Assesses contextual fit	Local-suitability assessment	Local farmer, advisor, and domain review	Seasonal and spatial variability remain

Field-validation planning	What evidence would future real-world assessment require?	Protocol, outcome measures, safeguards, stopping rules, and oversight	Structures future evaluation without claiming completion	Field-validation plan	Independent methodological and ethical review	A plan mistake complete
Advisory output review	Does the response remain within the approved output class?	Output taxonomy, wording controls, uncertainty, and source map	Detects recommendation-like or prohibited language	Approval, revision, escalation, or blocking	Reviewer and boundary-control assessment	Advice implied i
Human-in-the-loop review	Can a human materially change or stop the output before consequential use?	Reviewer authority, competence, evidence access, time, and independence	Approves, revises, rejects, or escalates	Documented review disposition	Rubber-stamping and automation bias	Mean over require
Agricultural advisor review where supported	Is professional agricultural interpretation required?	Local agronomic context and complete traceability package	Reviews validity and consequence	Advisor-reviewed interpretation or deferral	Qualified expert assessment	Profes review c guar correct
Expert escalation	Which unresolved cases require specialist attention?	High severity, evidence conflict, rarity, or local mismatch	Routes the case to an appropriate reviewer	Escalation record or deferred response	Incorrect routing or delayed review	Named e and re: crit
Farmer feedback	Does the output reflect observed conditions, constraints, and objectives?	Farmer observations and confirmation of assumptions	Detects contextual mismatch and supports correction	Confirmation, correction, rejection, or new evidence	Farmer participation and comprehension support	Feedbac incom influence sys
Output traceability	Can the complete generation and review process be reconstructed?	Inputs, prompt, retrieval, model version, output, checks, and reviewers	Enables verification and investigation	Trace package	Missing or inaccessible logs	Audi governan
Monitoring	Do output behaviour and control performance remain within acceptable bounds?	Drift indicators, grounding failures, deferrals, incidents, and feedback	Detects emerging risk	Alert, investigation, revision, or suspension	Human triage and governance review	Unrep failure remain

No validation-sufficiency boundary	Are validation findings restricted to their tested scope?	Context, sample, methods, limitations, and uncertainty	Prevents universal reliability claims	Bounded validation language	Independent methodological review	Reader overgeneralization, limited evidence
No deployment-readiness boundary	Is conceptual completeness distinguished from operational readiness?	Evidence of implementation, field testing, assurance, resourcing, and governance	Prevents premature deployment claims	Explicit conceptual-status statement	Independent readiness assessment would be required separately	Framework may be inappropriate for operational material

Safety Controls and Decision Boundaries

Safety principles become meaningful only when they are translated into operationally identifiable controls, assigned responsibilities, review stages, records, and escalation procedures. Applied AI-ethics methods commonly include impact assessment, documentation, auditing, stakeholder engagement, risk review, and lifecycle monitoring, although their maturity and effectiveness vary across settings [17]. For generative agricultural decision support, these mechanisms should begin with a clearly documented use case specifying the intended users, permitted data, allowable output classes, excluded decisions, expected review process, and foreseeable consequences of misuse. A conversational system designed to explain agricultural information should not gradually assume a recommendation or control function merely because users request increasingly specific answers. Use-case definition is therefore a continuing boundary rather than a one-time project description.

Risk thresholds determine how controls change as the potential consequence of an output increases. Informational content may require basic grounding and uncertainty disclosure, whereas recommendation-like content may require factual verification, agronomic review, local validation, human authorisation, and expert escalation. Ethical governance of AI requires practical mechanisms for verification, transparency, oversight, and institutional accountability rather than reliance on trust in designers or confidence in automated outputs [18]. In the proposed framework, high-risk query deferral is a safety function rather than a system failure. When evidence is insufficient, local context is incomplete, or potential consequences exceed the system's approved role, the appropriate output is a clarification request, restricted informational response, or referral to qualified human review.

Decision boundaries should be enforced at several points. At the interface level, the system should distinguish explanation from recommendation-like language and display material uncertainty prominently. At the generation level, it should restrict claims that lack source support or exceed retrieved evidence. At the workflow level, it should prevent unreviewed outputs from directly initiating machinery, irrigation, nutrient application, pest control, procurement, or other consequential actions. At the governance level, it should define who may approve, revise, reject, suspend, or escalate an output. General ethical principles cannot by themselves demonstrate that these controls are effective, consistently applied, or suitable for a particular agricultural environment [19]. The framework therefore includes explicit no-autonomous-action, no-hallucination-elimination, no-safety-guarantee, and no-deployment-readiness boundaries.

Safety controls should also be proportionate to consequence severity and should preserve transparency, contestability, and meaningful human authority. Algorithmic accountability depends on whether affected users can understand the basis of a decision, challenge its assumptions, and identify the actors responsible for its design and use [20]. Source traceability, response grounding, farmer confirmation, expert-review triggers, monitoring, and incident reporting can support these functions, but each remains fallible. Monitoring may fail to detect unreported harm; farmer confirmation may be influenced by persuasive system language; and expert review may be superficial when evidence is incomplete or

workloads are high. Feedback and knowledge-base updating should consequently be controlled, versioned, and verified so that unconfirmed corrections, anecdotal reactions, or model-generated claims do not become new sources of error.

Accountability in AI-Generated Farm Advice

Accountability begins with transparency about how farm data are collected, transformed, shared, and used. Farmers' trust in smart-farming services is affected by whether providers explain what happens to data, who benefits from their use, and how decisions are made [21]. In generative decision support, these concerns extend to the relationship between input data and output language. A farmer should be able to determine which observations, records, retrieved sources, and assumptions contributed to a consequential response. Farmer autonomy also requires the ability to question, reject, or correct an output without being treated as non-compliant with the system. Confirmation should therefore function as contextual verification rather than as a mechanism for transferring responsibility for system defects to the farmer.

Data stewardship responsibilities include access control, quality assurance, retention, provenance, consent, and appropriate reuse. Agricultural data may contain commercially sensitive information about production conditions, land, inputs, performance, and management practices. Unclear ownership, access, and privacy arrangements can reduce farmers' willingness to participate in digital systems and weaken the legitimacy of data-driven advice [22]. The proposed framework assigns a data-steward role for maintaining input records and a knowledge-base curator role for selecting, reviewing, versioning, and retiring agricultural sources. These roles may be performed by individuals or organisations, but their responsibilities should remain distinguishable. Data provenance does not prove accuracy, and knowledge-base curation does not eliminate outdated, contested, or locally unsuitable evidence.

Developer and platform accountability should likewise be separated. Developers are responsible for documenting model purpose, known limitations, evaluation conditions, version changes, and foreseeable failure modes. Platform providers control interfaces, integrations, access, monitoring, output restrictions, and escalation pathways. Agricultural advisors may contribute contextual or professional review where appropriate, but they should not be expected to endorse opaque outputs lacking traceable evidence. The social and ethical effects of agricultural AI extend across developers, technology providers, farmers, workers, advisors, communities, and institutions, making single-actor accountability inadequate [23]. A human review record should therefore identify what evidence was examined, what changes were made, why the output was approved or rejected, and which uncertainties remained unresolved.

Auditability requires records that allow a consequential output to be reconstructed, including relevant inputs, retrieval results, model and knowledge-base versions, generated claims, automated screening results, uncertainty disclosures, escalation decisions, and human interventions. Incident reporting should preserve these records and distinguish suspected hallucination, factual error, agronomic invalidity, boundary violation, privacy failure, cybersecurity event, and misuse. Platform concentration, unequal access, data control, and asymmetric bargaining power can shape which incidents are visible and whose interests determine corrective action [24]. Accountability must therefore include the capacity to investigate systemic patterns rather than attributing every failure to individual user judgement. The framework maps conceptual responsibilities but does not determine legal liability, prove regulatory compliance, or establish that named roles have sufficient authority, competence, or resources.

Proposed Safety Framework

The proposed framework treats hallucination risk as a systemic property of the relationship among agricultural data, knowledge resources, generative models, platform design, human interpretation, institutional responsibility, and farm-level consequence. Agricultural AI risks can emerge from data selection, model behaviour, optimisation goals, deployment arrangements, uneven adoption, and wider externalities, while responsible smart-farming innovation also requires attention to privacy, power, autonomy, inclusion, and the purposes served by technology [25-27]. The framework therefore does not position hallucination detection as an isolated software function. It organises a sequence of preventive,

interpretive, review, accountability, and learning stages through which generative agricultural outputs can be restricted or escalated according to their intended use and potential consequence.

The first stages establish the context within which any output may be generated. Agricultural use-case definition specifies permitted users, tasks, inputs, output classes, exclusions, and intended decision relationships. Stakeholder and responsibility mapping identifies developers, platform providers, data stewards, knowledge-base curators, agricultural reviewers, farmers, and relevant institutional actors. Advisory-risk classification then distinguishes information provision, explanation, interpretation, scenario analysis, recommendation-like content, and prohibited autonomous action. Data-quality and provenance review assesses whether sensor, soil, weather, crop, imagery, farm-record, and textual inputs are sufficiently complete, current, plausible, and traceable. Knowledge-base curation review evaluates source quality, currency, relevance, conflict, geographic fit, and version history before information is made available for retrieval.

The next stages focus on generation and validation. Retrieval and response-grounding review examines whether selected sources are relevant and whether generated claims remain faithful to them. Generative output remains a draft until hallucination-risk screening identifies unsupported statements, fabricated evidence, source-attribution errors, contradictions, overconfidence, and uncertainty underreporting. Factual accuracy checking verifies material claims independently of linguistic quality, while agronomic-validity review examines whether an interpretation is coherent for the relevant crop, environment, management setting, and local constraints. Uncertainty communication records missing data, conflicting evidence, plausible alternatives, and the limits of model and review processes. Decision-boundary enforcement then determines whether the output may be released as bounded information, must be revised, requires human review, or should be deferred.

Human review and expert escalation provide a further decision gate rather than an automatic approval mechanism. Reviewers require access to the complete evidence and traceability package, authority to alter or reject the response, and sufficient competence to interpret its intended use. Farmer feedback and confirmation can expose missing constraints, incorrect assumptions, or differences between system representations and observed conditions. Agricultural knowledge and advice increasingly operate through connected networks in which farmers, advisors, researchers, technology organisations, and data intermediaries have interacting roles [28]. Generative systems should therefore be embedded in these human knowledge networks rather than positioned as self-sufficient advisory authorities. Output traceability and audit logging record the complete pathway, while monitoring and incident reporting identify drift, recurring grounding failures, inappropriate confidence, escalation problems, or decision-boundary violations.

The final stages support controlled learning without implying that safety has been established. Verified corrections may inform knowledge-base updating, model or prompt review, interface revision, reviewer training, or changes to risk thresholds. Framework updating should occur after material system changes, incidents, new evidence, stakeholder feedback, or changes in the agricultural context. Every revision should preserve the distinction between conceptual safeguards and demonstrated operational performance. The framework contains explicit no-agronomic-advice, no-autonomous-decision, no-hallucination-elimination, no-safety-guarantee, and no-deployment-readiness boundaries. **Table 2** presents the proposed safety framework for managing hallucination risk in generative agricultural decision support. **Figure 1** presents the proposed safety framework for managing hallucination risk in generative agricultural decision support, linking advisory-risk classification, data grounding, validation, human review, decision boundaries, accountability, monitoring, and feedback updating.

Table 2. Proposed Safety Framework for Managing Hallucination Risk in Generative Agricultural Decision Support: Use-Case Definition, Responsibility Mapping, Advisory-Risk Classification, Data Provenance, Retrieval Grounding, Hallucination Screening, Factual Accuracy, Agronomic Validity, Uncertainty Communication, Decision Boundaries, Human Review, Farmer Feedback, Audit Logging, Monitoring, Incident Reporting, and Claim Boundaries

Framework stage	Core purpose	Required input	Safety function	Potential output	Validation or oversight	Failure risk
-----------------	--------------	----------------	-----------------	------------------	-------------------------	--------------

					need	
Agricultural use-case definition	Specify permitted tasks, users, contexts, and exclusions	Intended users, agricultural setting, data, decisions, and foreseeable consequences	Prevent ambiguous or expanding system scope	Approved use-case statement	Farmer, advisor, technical, and safety review	Scope creep or unintended reliance
Stakeholder and responsibility mapping	Identify actors affected by or responsible for the system	Developer, platform, data, knowledge, advisor, farmer, and institutional roles	Allocate control and review functions	Responsibility map	Multi-stakeholder review	Unowned controls or conflicting authority
Advisory-risk classification	Match output class to potential consequence	Intended output, severity, reversibility, uncertainty, and affected resources	Determine required safeguards	Informational, restricted, escalated, or prohibited class	Domain and safety review	Underclassification
Data quality and provenance review	Assess whether inputs are reliable and reconstructable	Sensor, weather, soil, crop, imagery, record, and metadata inputs	Reject incomplete, implausible, stale, or untraceable data	Accepted, qualified, or rejected input set	Data steward and domain oversight	Valid-looking but misleading data
Knowledge-base curation review	Maintain relevant and traceable agricultural evidence	Candidate sources, dates, geographic scope, conflicts, and versions	Control source inclusion, updating, and retirement	Versioned agricultural knowledge base	Agricultural and evidence-review expertise	Stale, biased, conflicting, or incomplete content
Retrieval-quality review	Assess whether retrieved evidence fits the query and context	Query, ranking results, source metadata, and relevance criteria	Detect irrelevant or weak retrieval	Accepted or revised retrieval set	Technical and domain sampling	Superficially similar but unsuitable sources
Response-grounding review	Ensure generated claims remain faithful to evidence	Retrieved passages, input data, and generated claim map	Detect unsupported extension or source distortion	Grounded, revised, or blocked claim set	Automated checks and human review	Nuanced mismatch may evade detection
Generative output generation	Produce bounded explanatory or	Approved inputs, sources,	Restrict generation to	Draft output	Automated controls and later review	Recommendation leakage or

[illegible]

Decision-boundary enforcement	Restrict outputs according to approved authority	Risk class, prohibited language, action pathways, and evidence status	Block advisory or autonomous-action leakage	Release, revision, escalation, or deferral	Product and safety governance	Indirect wording may bypass restrictions
Human review	Preserve meaningful human authority over consequential output	Trace package, reviewer competence, authority, and review criteria	Approve, revise, reject, or suspend	Documented review disposition	Competent and independent review	Automation bias or rubber-stamping
Expert escalation	Route unresolved or high-consequence cases	Risk triggers, evidence conflict, rarity, and local mismatch	Obtain specialist interpretation	Expert review request or deferred response	Named reviewer role and response process	Incorrect routing or unavailable expertise
Farmer feedback and confirmation	Verify assumptions, constraints, observations, and intended use	Farmer knowledge, field observations, objectives, and accessible output summary	Preserve autonomy and detect local mismatch	Confirmation, correction, rejection, or escalation	Voluntary, understandable participation	Persuasive language may influence confirmation
Output traceability and audit logging	Reconstruct how each consequential output was produced and reviewed	Inputs, prompts, retrieval, versions, output, checks, and human actions	Support verification and investigation	Traceable output record	Access, retention, and audit governance	Missing, altered, or inaccessible records
Monitoring and drift review	Detect changing behaviour and control performance	Grounding errors, deferrals, disagreements, version changes, and user reports	Identify model or knowledge-base drift	Alert, investigation, revalidation, or suspension	Human triage and governance review	Hidden or unreported failure
Incident reporting	Preserve and investigate suspected misleading or harmful outputs	User report, trace record, severity, and system context	Support corrective action and organisational learning	Incident record and response	Independent or cross-functional investigation	Underreporting or blame shifting

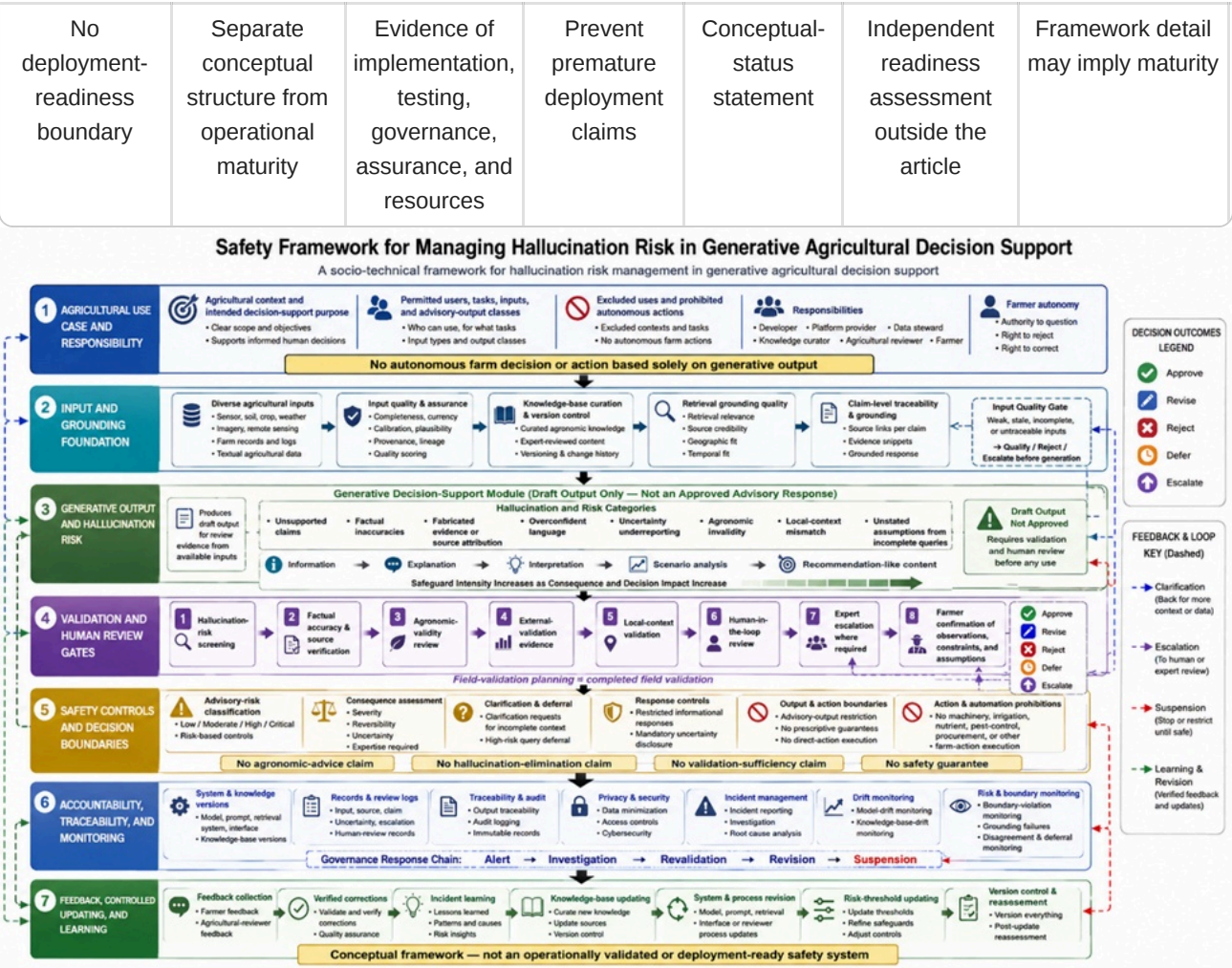


Figure 1. Safety Framework for Managing Hallucination Risk in Generative Agricultural Decision Support

Alt-text description: A conceptual safety framework showing generative agricultural decision-support outputs moving through data grounding, hallucination screening, factual accuracy checks, agronomic validity review, uncertainty communication, decision boundaries, human review, expert escalation, farmer feedback, accountability, monitoring, and updating loops.

Practical Implications and Future Research

For farmers, agricultural advisors, extension services, cooperatives, and other knowledge intermediaries, the framework highlights the need to distinguish conversational accessibility from evidentiary reliability. Users require clear disclosure of data sources, assumptions, uncertainty, output classification, and the limits of system authority. Local validation and expert escalation are particularly important where an apparently general answer may be affected by crop variety, soil conditions, climate, production system, management history, or farmer objectives. Future studies should examine how different users interpret uncertainty language, traceability displays, deferral messages, and distinctions between information and recommendation-like content. Research should also assess whether human-review designs genuinely improve challenge and correction or instead produce automation bias and administrative sign-off.

For developers, platform providers, data stewards, and knowledge-base curators, priority areas include agricultural hallucination evaluation, claim-level grounding assessment, source-verification methods, curated knowledge governance,

model and retrieval drift monitoring, and privacy- and cybersecurity-aware traceability. Retrieval-supported generation should be evaluated as a multi-stage process in which source selection, ranking, contextual fit, evidence interpretation, and language generation may fail independently [8]. Longitudinal research is needed to determine how model updates, seasonal changes, knowledge-base revisions, and shifting agricultural conditions affect previously assessed controls. Incident-reporting research should investigate practical taxonomies, reporting incentives, evidence preservation, independent review, and mechanisms for converting verified incidents into controlled system and framework updates.

For researchers and policymakers, the framework identifies a need for interdisciplinary work linking technical evaluation with agronomy, agricultural knowledge systems, human factors, ethics, governance, and rural digital inclusion. Local and experiential knowledge should be studied as part of validation rather than as an informal supplement to supposedly complete digital information [15]. Accountability research should examine whether developers, platforms, advisors, cooperatives, data actors, and public institutions possess sufficient authority and resources to perform their assigned responsibilities, while farmer-trust research should consider transparency, data use, benefit distribution, and meaningful autonomy [21]. These implications are conceptual research directions rather than agronomic advice, legal advice, policy prescriptions, procurement recommendations, platform-selection guidance, or deployment instructions.

Conclusion

This article contributes a conceptual AI safety framework for managing hallucination risk in generative agricultural decision support by linking risk interpretation, advisory-output classification, farm-level consequence assessment, data and knowledge grounding, factual verification, agronomic-validity review, uncertainty communication, human oversight, expert escalation, farmer feedback, decision boundaries, accountability, monitoring, incident reporting, and controlled updating. The framework treats hallucination as one element of a broader socio-technical safety problem that also includes factual inaccuracy, unsupported output, fabricated evidence, local-context mismatch, weak validation, overconfidence, and responsibility gaps. Its central implication is that agricultural generative systems should not be evaluated through conversational fluency or apparent model confidence. Safer decision-support workflows require traceable evidence, proportionate validation, meaningful human authority, explicit uncertainty, accountable roles, and enforceable restrictions on advice-like or autonomous action. These controls may help organise risk management, but they do not eliminate hallucination, guarantee agronomic correctness, establish legal compliance, certify operational safety, or demonstrate deployment readiness.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 30 Jul 2025

Revised: 12 Sep 2025

Accepted: 06 Nov 2025

Published online: 25 May 2026

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

- Wolfert S, Ge L, Verdouw C, Bogaardt MJ. Big data in smart farming—A review. *Agric Syst*. 2017;153:69-80.
<https://doi.org/10.1016/j.agsy.2017.01.023>.
- Klerkx L, Jakku E, Labarthe P. A review of social science on digital agriculture, smart farming and agriculture 4.0: New contributions and a future research agenda. *NJAS Wageningen J Life Sci*. 2019;90-91:100315.
<https://doi.org/10.1016/j.njas.2019.100315>.
- Liakos KG, Busato P, Moshou D, Pearson S, Bochtis D. Machine learning in agriculture: A review. *Sensors (Basel)*. 2018;18(8):2674.
<https://doi.org/10.3390/s18082674>.
- Dara R, Hazrati Fard SM, Kaur J. Recommendations for ethical and responsible use of artificial intelligence in digital agriculture. *Front Artif Intell*. 2022;5:884192.
<https://doi.org/10.3389/frai.2022.884192>.
- Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. *ACM Comput Surv*. 2023;55(12):248.
<https://doi.org/10.1145/3571730>.
- Alkaiissi H, McFarlane SI. Artificial hallucinations in ChatGPT: Implications in scientific writing. *Cureus*. 2023;15(2):e35179.
<https://doi.org/10.7759/cureus.35179>.
- Hicks MT, Humphries J, Slater J. ChatGPT is bullshit. *Ethics Inf Technol*. 2024;26:38.
<https://doi.org/10.1007/s10676-024-09775-5>.
- Zhang W, Zhang J. Hallucination mitigation for retrieval-augmented large language models: A review. *Mathematics*. 2025;13(5):856.
<https://doi.org/10.3390/math13050856>.
- Kuska MT, Wahabzada M, Paulus S. AI for crop production—Where can large language models provide substantial value? *Comput Electron Agric*. 2024;221:108924.
<https://doi.org/10.1016/j.compag.2024.108924>.
- Shaikh TA, Rasool T, Mir WA. Fields of the future: Digital transformation in smart agriculture with large language models and generative AI. *Comput Stand Interfaces*. 2025;94:104005.
<https://doi.org/10.1016/j.csi.2025.104005>.
- Li H, Wu H, Li Q, Zhao C. A review on enhancing agricultural intelligence with large language models. *Artif Intell Agric*. 2025;15(4):671-85.
<https://doi.org/10.1016/j.aiia.2025.05.006>.

Chen D, Huang Y. Integrating reinforcement learning and large language models for crop production process management optimization and control through a new knowledge-based deep learning paradigm. *Comput Electron Agric.* 2025;232:110028. <https://doi.org/10.1016/j.compag.2025.110028>.

Shneiderman B. Human-centered artificial intelligence: Reliable, safe and trustworthy. *Int J Hum Comput Interact.* 2020;36(6):495-504. <https://doi.org/10.1080/10447318.2020.1741118>.

Cartolano A, Cuzzocrea A, Pilato G. Analyzing and assessing explainable AI models for smart agriculture environments. *Multimed Tools Appl.* 2024;83(12):37225-46. <https://doi.org/10.1007/s11042-023-17978-z>.

Ingram J, Maye D. What are the implications of digitalisation for agricultural knowledge?. *Front Sustain Food Syst.* 2020;4:66. <https://doi.org/10.3389/fsufs.2020.00066>.

Eastwood C, Klerkx L, Ayre M, Dela Rue B. Managing socio-ethical challenges in the development of smart farming: From a fragmented to a comprehensive approach for responsible research and innovation. *J Agric Environ Ethics.* 2019;32(5-6):741-68. <https://doi.org/10.1007/s10806-017-9704-5>.

Morley J, Floridi L, Kinsey L, Elhalal A. From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Sci Eng Ethics.* 2020;26(4):2141-68. <https://doi.org/10.1007/s11948-019-00165-5>.

Winfield AFT, Jirotko M. Ethical governance is essential to building trust in robotics and artificial intelligence systems. *Philos Trans A Math Phys Eng Sci.* 2018;376(2133):20180085. <https://doi.org/10.1098/rsta.2018.0085>.

Mittelstadt B. Principles alone cannot guarantee ethical AI. *Nat Mach Intell.* 2019;1(11):501-7. <https://doi.org/10.1038/s42256-019-0114-4>.

Lepri B, Oliver N, Letouzé E, Pentland A, Vinck P. Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges. *Philos Technol.* 2018;31(4):611-27. <https://doi.org/10.1007/s13347-017-0279-x>.

Jakku E, Taylor B, Fleming A, Mason C, Fielke S, Sounness C, et al. "If they don't tell us what they do with it, why would we trust them?" Trust, transparency and benefit-sharing in smart farming. *NJAS Wageningen J Life Sci.* 2019;90-91:100285. <https://doi.org/10.1016/j.njas.2018.11.002>.

Wiseman L, Sanderson J, Zhang A, Jakku E. Farmers and their data: An examination of farmers' reluctance to share their data through the lens of the laws impacting smart farming. *NJAS Wageningen J Life Sci.* 2019;90-91:100301. <https://doi.org/10.1016/j.njas.2019.04.007>.

Ryan M. The social and ethical impacts of artificial intelligence in agriculture: Mapping the agricultural AI literature. *AI Soc.* 2023;38(6):2473-85. <https://doi.org/10.1007/s00146-021-01377-9>.

Rotz S, Duncan E, Small M, Botschner J, Dara R, Mosby I, et al. The politics of digital agricultural technologies: A preliminary review. *Sociol Ruralis.* 2019;59(2):203-29. <https://doi.org/10.1111/soru.12233>.

Tzachor A, Devare M, King B, Avin S, Ó hÉigeartaigh S. Responsible artificial intelligence in agriculture requires systemic understanding of risks and externalities. *Nat Mach Intell.* 2022;4(2):104-9. <https://doi.org/10.1038/s42256-022-00440-4>.

van der Burg S, Bogaardt MJ, Wolfert S. Ethics of smart farming: Current questions and directions for responsible innovation towards the future. *NJAS Wageningen J Life Sci.* 2019;90-91:100289.

<https://doi.org/10.1016/j.njas.2019.01.001>.

Rose DC, Chilvers J. Agriculture 4.0: Broadening responsible innovation in an era of smart farming. *Front Sustain Food Syst.* 2018;2:87.

<https://doi.org/10.3389/fsufs.2018.00087>.

Fielke S, Taylor B, Jakku E. Digitalisation of agricultural knowledge and advice networks: A state-of-the-art review. *Agric Syst.* 2020;180:102763.

<https://doi.org/10.1016/j.agry.2019.102763>.