

ORIGINAL RESEARCH

Open access

Say What the Model Does Not Know: A Framework for Confidence Calibration, Ambiguity Disclosure, Risk Communication, and Selective Prediction in Explainable Decision Intelligence

Maria Lopez^{1*}, David Thompson²

Abstract

Intelligent systems increasingly provide predictions, recommendations, and explanations in decisions where uncertainty can materially affect interpretation and action. Yet many explainable systems foreground predicted outcomes or feature contributions while inadequately communicating what the model does not know. This article proposes a conceptual uncertainty-aware explainable decision-intelligence framework organized around four connected responsibilities: calibrating confidence, disclosing ambiguity, communicating decision-relevant risk, and supporting selective prediction or abstention when available evidence is insufficient. The framework treats uncertainty as a trustworthiness requirement rather than a secondary technical output. It distinguishes aleatoric uncertainty arising from irreducible variability, epistemic uncertainty associated with limited knowledge or evidence, and distributional uncertainty arising when operating conditions depart from those represented during development. It further connects confidence calibration and miscalibration review with prediction qualification, rejection, abstention, deferral, and human review. Ambiguity disclosure and risk communication are positioned as interpretive functions that should clarify scope limitations, competing interpretations, possible consequences, and appropriate next actions without implying that communication alone guarantees understanding. Human decision support is addressed through escalation pathways, oversight, trust calibration, and evaluation of overreliance, underreliance, automation bias, and cognitive burden. The principal contribution is an integrated framework that links uncertainty identification, calibrated interpretation, selective system behavior, human review, monitoring, and feedback updating. The framework supports conceptual development and domain-specific evaluation but does not independently establish correctness, safety, decision validity, ethical sufficiency, legal compliance, or deployment readiness.

Keywords Explainable AI, Decision intelligence, Uncertainty-aware AI, Confidence calibration, Selective prediction, Risk communication

*Correspondence:

Maria Lopez
maria.lopez@ie.edu

¹ Department of AI and Decision Intelligence, IE University, Madrid, Spain

² Department of Computer Science, University of Auckland, Auckland, New Zealand

Introduction

Explainable artificial intelligence has expanded the vocabulary through which intelligent systems communicate predictions, influential features, decision logic, and model behavior. However, an explanation of why an output was produced is incomplete when it omits whether the supporting evidence is weak, the input is unfamiliar, the label is

ambiguous, or the explanation itself is unstable. Explainable decision intelligence therefore requires more than output attribution: explanations must be intelligible, sensitive to their recipients, and bounded by the limitations of the underlying model, especially in consequential settings [1-3].

The practical concern is not simply whether users trust an intelligent system, but whether their reliance corresponds to the system's actual and context-specific capabilities. Trust may be shaped by perceived accuracy, transparency, predictability, embodiment, prior expectations, and organizational context. A confident prediction or persuasive explanation can consequently encourage reliance even when the system is operating beyond its validated scope. Trustworthiness therefore requires calibrated expectations rather than undifferentiated confidence in artificial intelligence [4].

An uncertainty-aware perspective asks intelligent systems to communicate what is not established as clearly as what is predicted. Such communication may include qualified confidence, alternative interpretations, evidence limitations, distribution-shift warnings, requests for additional information, or explicit refusal to produce an unsupported answer. These responses do not make uncertainty disappear. Their purpose is to prevent prediction interfaces from converting incomplete evidence into an appearance of certainty and to support decisions that remain sensitive to consequence, scope, and residual risk.

This article develops an original conceptual framework for uncertainty-aware explainable decision intelligence. The framework integrates four responsibilities: confidence calibration, ambiguity disclosure, risk communication, and selective prediction or abstention. These responsibilities are connected to uncertainty-source identification, human review, escalation, monitoring, and feedback updating. The contribution is organizational rather than prescriptive: it structures the questions that designers and evaluators should examine without selecting universal estimators, calibration methods, visualization formats, decision thresholds, or deployment procedures.

Uncertainty as a Trustworthiness Requirement

Trustworthy artificial intelligence is commonly understood as a multidimensional objective involving technical performance, transparency, robustness, fairness, accountability, privacy, and human agency. Within this broader structure, uncertainty is a trustworthiness requirement because predictions cannot be interpreted responsibly without information about evidential strength, model limitations, and operating scope. Nevertheless, uncertainty awareness is only one component of trustworthiness and cannot substitute for independent assessment of other technical and sociotechnical properties [5].

Research on uncertainty quantification demonstrates that uncertainty may originate from data variability, model knowledge, approximation choices, distributional mismatch, and interactions among these sources. The literature also shows substantial diversity in estimation assumptions, computational techniques, applications, and evaluation practices. No single uncertainty approach establishes trustworthy operation across all tasks, populations, models, or environments, and uncertainty estimates must be interpreted within the conditions under which they were produced [6-8].

A system may be accurate on average yet misleadingly confident in particular regions, subgroups, or unfamiliar cases. Conversely, it may be systematically underconfident and reject useful predictions unnecessarily. Uncertainty-aware decision intelligence must therefore distinguish model-generated confidence from validated predictive confidence and must examine whether confidence retains its meaning under relevant decision conditions. Uncertainty quantification can support such assessment, but it does not independently verify that a prediction is correct, fair, robust, causally valid, or suitable for action [6].

Risk-sensitive decisions further require uncertainty to be connected with consequences. The same degree of uncertainty may be tolerable in a reversible, low-consequence recommendation but unacceptable where errors are difficult to detect, costly to reverse, or unequally distributed. An uncertainty-aware system should therefore communicate limitations in relation to the decision being supported and should define conditions under which prediction, qualification, abstention, or

escalation may be considered. This requirement establishes a no-uncertainty-guarantee boundary and a no-deployment-readiness boundary: uncertainty information may improve interpretability, but it is not evidence that the full system is ready for operational use [5].

Aleatoric, Epistemic, and Distributional Uncertainty

Aleatoric uncertainty concerns variability that is inherent in observations, outcomes, or the process being modeled, whereas epistemic uncertainty concerns limitations in the model's knowledge arising from incomplete evidence, restricted coverage, or uncertain assumptions. The distinction matters because the two forms of uncertainty imply different responses. Aleatoric uncertainty may require prediction ranges or outcome qualification, while epistemic uncertainty may motivate additional evidence, model revision, human review, or abstention. The categories are analytically useful but cannot by themselves determine an appropriate action [9].

Predictive uncertainty combines multiple sources into an assessment of uncertainty surrounding a particular output. Empirical comparisons show that the quality of such assessments is dependent on the estimation method, model architecture, task, data-generating conditions, and evaluation procedure. A method that produces informative uncertainty in one prediction context may not retain the same properties in another. Consequently, uncertainty estimates should not be transferred across domains or model classes without renewed evaluation [10].

Implementation constraints introduce further complications. Scalable uncertainty estimators may differ in computational cost, calibration behavior, sensitivity to model misspecification, and capacity to identify unfamiliar cases. Comparative evidence indicates that uncertainty methods can produce materially different results even when applied to the same broad prediction problem. The framework therefore treats estimator selection as a domain-specific methodological decision rather than recommending one technique as universally superior [11].

Domain expertise may also be needed to determine whether an uncertainty signal corresponds to a meaningful ambiguity. In diagnostic decision support, for example, expert assessment can help evaluate whether uncertain predictions align with difficult, atypical, or clinically ambiguous cases. Such validation does not prove that uncertainty estimates will remain useful under every population or deployment condition, but it illustrates why technical uncertainty assessment should be connected with domain interpretation and human review [12].

Distributional uncertainty arises when current inputs, outcomes, or relationships differ from those represented during model development. It includes dataset shift, out-of-distribution conditions, and novelty, while unknown unknowns represent residual failure possibilities that may not be captured by any predefined uncertainty category. Data uncertainty, label disagreement, measurement error, model uncertainty, explanation instability, and predictive uncertainty may overlap in practice. The framework consequently includes both a no-uncertainty-guarantee boundary and a no-uncertainty-type-sufficiency boundary: identifying a category helps structure investigation but does not establish decision validity. **Table 1** summarises uncertainty types and sources that must be distinguished before uncertainty-aware explanations can support trustworthy decision intelligence.

Table 1. Aleatoric, Epistemic, and Distributional Uncertainty in Explainable Decision Intelligence: Data Uncertainty, Label Uncertainty, Measurement Uncertainty, Model Uncertainty, Explanation Uncertainty, Prediction Uncertainty, Ambiguity, Unknown Unknowns, Dataset Shift, Out-of-Distribution Conditions, Novelty, and Uncertainty Boundaries

Uncertainty type or source	Core trustworthiness question	What the uncertainty indicates	What it cannot establish	Potential decision risk	Communication need	Evaluation need
Aleatoric uncertainty	How much irreducible	Inherent noise or outcome	Model correctness	False precision in variable	Explain unavoidable	Compare estimates with

	variability affects the outcome?	variability	or suitability	cases	variability	observed variation
Epistemic uncertainty	Where is knowledge or evidence limited?	Reducible model or evidence limitation	That additional data resolves every error	Overconfidence in sparse regions	Identify knowledge gaps	Sensitivity and external validation
Distributional uncertainty	Does the case match evaluated conditions?	Possible operating-distribution mismatch	That every shifted case is erroneous	Confident out-of-scope prediction	Provide scope or shift warning	Test realistic shifts
Data uncertainty	Are inputs incomplete, noisy, or unrepresentative?	Data-quality limitations	Total system uncertainty	Decisions based on defective evidence	Disclose quality and provenance limitations	Data-quality assessment
Label uncertainty	Is the target contested or inconsistently assigned?	Annotation disagreement or ambiguity	A uniquely correct label	Artificial consensus	Preserve relevant disagreement	Agreement and subgroup analysis
Measurement uncertainty	How reliable are observed quantities?	Measurement error or variability	Model or decision validity	Incorrect threshold crossing	Communicate measurement limitations	Instrument and propagation assessment
Model uncertainty	How dependent is the output on model assumptions?	Parameter or structural uncertainty	Real-world causal validity	Fragile predictions presented as stable	Qualify model dependence	Model comparison and perturbation analysis
Explanation uncertainty	How stable and complete is the explanation?	Variability or incompleteness in explanatory accounts	Prediction correctness	Unstable explanation treated as fact	Distinguish explanation from prediction uncertainty	Fidelity and stability evaluation
Prediction uncertainty	How uncertain is the specific output?	Uncertainty surrounding a case-level prediction	The source of uncertainty	Misuse of a point estimate	Present outcome qualification	Calibration and task validation
Ambiguity	Are several interpretations defensible?	Non-uniqueness or underspecification	Which interpretation is normatively correct	Unsupported definitive output	State plausible alternatives	Comprehensiveness and usefulness testing
Unknown unknowns	Could unrecognized failures remain?	Residual uncertainty beyond identified categories	Probability or form of every future failure	Undetected novel harm	State residual risk	Stress testing and incident learning

Dataset shift	Have population or relationships changed?	Change in operational conditions	Exact performance loss	Silent degradation	Communicate affected scope	Continuous monitoring
Out-of-distribution condition	Is the input unfamiliar relative to development data?	Possible scope violation	Definite error	Unsupported classification	Provide out-of-scope warning	OOD evaluation
Novelty	Does the case contain a previously unmodeled pattern?	Departure from known normality	Harmfulness of novelty	Forced treatment as a known case	State novelty and uncertainty	Domain review
No uncertainty-guarantee boundary	Which assurances remain unestablished?	Limits of uncertainty evidence	Correctness, safety, legality, or readiness	Uncertainty treated as certification	Explicit non-guarantee statement	Independent property evaluation
No uncertainty-type sufficiency boundary	Is classifying uncertainty sufficient for action?	Structured uncertainty diagnosis	Appropriate threshold or decision	Taxonomy converted into prescription	Explain that categories guide assessment	Domain-specific consequence testing

Confidence Calibration and Selective Prediction

Confidence calibration examines whether stated predictive probabilities correspond to observed outcome frequencies under relevant conditions. Selective prediction extends this concern from interpretation to system behavior by asking whether the system should issue, qualify, reject, or defer a prediction. Together, calibration assessment, rejection options, and uncertainty-informed model use create a pathway from confidence evidence to conditional action, but none establishes operational readiness independently [13-15].

Selective classification may retain cases judged sufficiently supported while rejecting others, but confidence alone may be inadequate when inputs are out of distribution. A system can assign high softmax confidence to unfamiliar cases even when those cases differ substantially from its development environment. Distributional checks should therefore complement confidence assessment when defining prediction boundaries, particularly where out-of-scope cases could receive apparently decisive outputs [16].

A practical pathway begins with uncertainty-source assessment, followed by calibration review and identification of miscalibration, overconfidence, underconfidence, low-confidence cases, and high-uncertainty cases. The system may then issue an ordinary prediction, provide a qualified response, request additional evidence, abstain, or defer the case for human review. These alternatives define a prediction boundary rather than a binary assumption that every input must receive an answer. **Figure 1** illustrates the confidence calibration and selective prediction pathway, showing how uncertainty assessment, calibration review, confidence qualification, abstention, deferral, and human review can shape decision-boundary behavior.

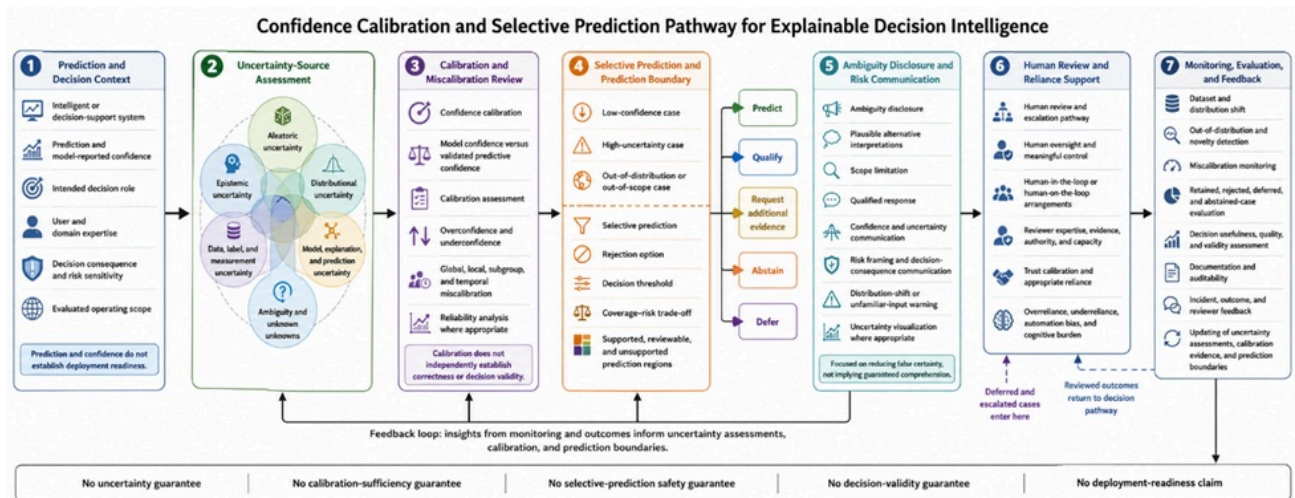


Figure 1. Confidence Calibration and Selective Prediction Pathway for Explainable Decision Intelligence

Alt-text description:

A conceptual pathway showing model prediction moving through uncertainty assessment, confidence calibration, miscalibration review, selective prediction, abstention, deferral, human review, risk communication, and decision boundaries.

Calibration assessment should not be reduced to one aggregate statistic. Reliability diagrams, calibration errors, subgroup analyses, and local error inspection may reveal different forms of miscalibration, while each method introduces assumptions concerning binning, sample size, aggregation, and representativeness. A model may appear calibrated overall while remaining overconfident for consequential subgroups or boundary cases. The no-calibration-sufficiency boundary therefore requires calibration to be interpreted alongside discrimination, robustness, distributional applicability, decision consequences, and human-use conditions [13].

Selective prediction similarly redistributes rather than eliminates risk. Abstention can reduce unsupported answers, but rejected cases still require an appropriate downstream pathway. Deferral may fail when reviewers lack expertise, capacity, authority, or additional evidence, and prediction withholding may produce delay or unequal access. Coverage–risk trade-offs and decision thresholds must therefore be assessed in relation to retained errors, excluded cases, escalation quality, and consequences. The no-selective-prediction-safety-guarantee boundary makes clear that selective prediction can help structure safer decision boundaries without proving that the resulting system is safe [14].

Ambiguity Disclosure and Risk Communication

Uncertainty communication should identify what is uncertain, why it matters, and what the recipient may reasonably do next. Visual uncertainty representations can support interpretation, but their effectiveness depends on the judgment task, evaluation criteria, display design, and user population. A technically faithful display may still fail when it obscures the relevant comparison, imposes excessive cognitive demand, or encourages an unintended heuristic [17].

Confidence communication must therefore provide meaning rather than merely display a score. Users need to know whether confidence refers to a model-internal preference, an empirically calibrated probability, an interval, or another uncertainty representation. Evidence from AI-assisted decision-making indicates that the influence of uncertainty presentation varies with users' initial judgments, characteristics, and decision conditions. The same presentation may promote appropriate reliance for some users while producing confusion, overcorrection, or underreliance for others [18].

Ambiguity disclosure addresses cases in which several labels, interpretations, explanations, or courses of action remain defensible. A system may state that evidence is incomplete, alternatives remain plausible, or the case falls near a decision boundary. However, numerical uncertainty without action-oriented guidance may leave users aware that the system is unsure but unable to determine whether they should gather evidence, consult an expert, defer action, or disregard the output. Capability-focused guidance may therefore complement uncertainty representation by clarifying what the system can and cannot support [19].

Risk communication connects uncertainty with decision consequences. It can distinguish the probability of error from the seriousness, reversibility, and distribution of possible outcomes. Uncertainty visualization may support this process by presenting ranges, distributions, or alternative outcomes, but display format can influence reasoning strategies and effect-size judgments. Evaluation should therefore examine not only whether users notice uncertainty, but also whether they interpret it consistently with the intended decision question [20].

The framework treats ambiguity disclosure, scope limitations, qualified responses, out-of-scope warnings, and distribution-shift warnings as related but distinct communication functions. Their purpose is to reduce false certainty without replacing human judgment or independent validation. The no-ambiguity-disclosure-sufficiency boundary recognises that visible limitations do not guarantee comprehension, while the no-risk-communication-guarantee boundary recognises that understanding a risk does not ensure an appropriate response. Communication must be evaluated through task-specific comprehension, reasoning, reliance, and decision consequences [17].

Uncertainty-Aware Human Decision Support

Human involvement should not be treated as a universal remedy for uncertain automation. Algorithm-in-the-loop decision processes can reproduce institutional assumptions, constrain the information available to decision-makers, and encourage nominal rather than substantive oversight. Human review contributes only when reviewers possess relevant expertise, sufficient time, meaningful authority, and access to evidence that can improve on the system's unsupported or uncertain output [21].

Interaction design may influence whether reviewers engage in independent reasoning. Cognitive forcing functions, such as requiring an initial judgment before revealing AI advice or prompting users to explain disagreement, can reduce overreliance in some settings. Nevertheless, added friction may increase workload, delay decisions, or be bypassed under pressure. Such mechanisms should therefore be assessed as context-dependent design options rather than universal safeguards [22].

Trust calibration concerns whether users' beliefs about system capability correspond to actual capability, whereas appropriate reliance concerns whether users accept useful advice and reject incorrect or unsupported advice. These constructs should be evaluated behaviorally and over time. Self-reported trust alone cannot reveal whether users systematically follow high-confidence errors, disregard correct recommendations, or change their reliance when the system operates outside familiar conditions [23].

Human intuition and domain knowledge further shape how explanations and uncertainty are interpreted. Users may reject correct AI advice because it conflicts with prior expectations, or accept incorrect advice because an explanation appears plausible. Domain experts and nonexperts may also require different levels of detail, uncertainty representation, and escalation support. Human-centered assessment should therefore examine how explanation type, expertise, intuition, and decision context interact with reliance [24].

The proposed framework distinguishes human review, human oversight, human control, human-in-the-loop operation, and human-on-the-loop supervision. It also requires an explicit escalation pathway that identifies who receives an uncertain case, what information accompanies it, and how the outcome returns to monitoring and feedback processes. Human participation does not independently establish decision quality or validity, and calibrated reliance does not guarantee

beneficial outcomes. **Table 2** maps confidence calibration, selective prediction, ambiguity disclosure, risk communication, and human decision-support requirements for uncertainty-aware explainable systems.

Table 2. Confidence Calibration, Selective Prediction, Ambiguity Disclosure, Risk Communication, and Human Decision-Support Requirements: Miscalibration, Overconfidence, Underconfidence, Abstention, Deferral, Prediction Boundaries, Qualified Responses, Uncertainty Visualization, Trust Calibration, Reliance, Automation Bias, and Decision-Validity Boundaries

Uncertainty-aware requirement	Core decision or communication question	Required evidence, assessment, or practice	Decision-support function	Potential output	Failure or misinterpretation risk	Human review or evaluation need
Confidence calibration	Does expressed confidence correspond to outcomes?	Representative outcome-linked assessment	Qualify confidence interpretation	Calibrated confidence statement	Calibration treated as correctness	Local, subgroup, and temporal evaluation
Miscalibration review	Where is confidence misleading?	Overconfidence and underconfidence analysis	Identify unreliable regions	Miscalibration warning	Aggregate measures hide local failure	Inspect consequential errors
Selective prediction	Should the system predict this case?	Uncertainty, scope, and consequence assessment	Restrict unsupported outputs	Predict or withhold	Rejection assumed to remove all risk	Evaluate retained and excluded cases
Abstention	Should the system decline to answer?	Evidence-insufficiency criteria	Prevent forced prediction	Qualified abstention	No useful next step	Assess downstream handling
Deferral	Should the case pass to another decision-maker?	Reviewer capability and pathway design	Transfer responsibility	Escalation request	Reviewer shares the same limitations	Review quality and capacity assessment
Prediction boundary	Which cases are supported, reviewable, or out of scope?	Calibration, shift, uncertainty, and consequence evidence	Delimit system authority	Predict, qualify, defer, or abstain	Brittle or untested boundaries	Boundary-capability testing
Ambiguity disclosure	What interpretations remain plausible?	Identification of alternatives and disagreement	Prevent false singularity	Alternative interpretations	Confusion or responsibility shifting	Comprehensive evaluation
Qualified response	How should the output be limited?	Evidence-linked caveats	Avoid absolute assertion	Conditional answer	Qualification overlooked	Salience and usability testing
Risk communication	What could follow from	Likelihood and consequence analysis	Connect uncertainty with action	Risk summary	Alarmism or false reassurance	Behavioral and comprehensive testing

	acting on the output?					
Uncertainty visualization	Which display supports the intended judgment?	Task-sensitive visual evaluation	Make ranges or distributions perceptible	Visual uncertainty representation	Bias, clutter, or anchoring	Perceptual and decision testing
Human review	Can a reviewer contribute independent evidence or judgment?	Expertise, authority, time, and information	Reassess uncertain cases	Review recommendation	Rubber-stamping or fatigue	Measure reasoning and outcomes
Trust calibration	Do user expectations match system capability?	Capability-specific belief assessment	Align expectations	Trust profile	Global trust despite variable performance	Longitudinal evaluation
Appropriate reliance	Do users accept helpful and reject unsupported advice?	Behavioral reliance measurement	Coordinate human and AI input	Acceptance or override	Overreliance or underreliance	Include correct and incorrect cases
Automation bias	Does AI advice suppress independent judgment?	Independent-decision and error analysis	Identify cognitive dependence	Bias warning or interaction change	Symbolic human oversight	Evaluate omission and commission errors
Cognitive load	Can users process uncertainty within task constraints?	Workload and comprehension assessment	Maintain usability	Adapted representation	Information overload	User- and expertise-specific testing
Decision-validity boundary	Is the final decision justified for its substantive purpose?	Independent domain standards	Separate prediction from valid action	Validity limitation	Technical performance treated as authorization	Domain-specific assessment
No human-review sufficiency boundary	Does human presence guarantee improvement?	Review-process evidence	Prevent oversight overclaiming	Non-sufficiency statement	Human involvement used as assurance	Demonstrate actual contribution
No reliance-benefit guarantee boundary	Does calibrated reliance guarantee better outcomes?	Outcome and consequence evaluation	Prevent behavioral proxy claims	Residual-risk statement	Reliance detached from consequences	Assess final decisions and harms

Proposed Explainable Decision-Intelligence Framework

The framework begins by defining the use case, prediction task, intended decision role, relevant users, operating environment, and potential consequences. It then identifies uncertainty sources before selecting a communication or action pathway. Novelty and anomaly detection are relevant at this stage because unfamiliar cases may not be adequately represented by ordinary confidence scores. Recognising departure from learned normality can help identify candidates for qualification, withholding, or review without implying that every novel case is erroneous [25].

The second stage distinguishes aleatoric, epistemic, and distributional uncertainty and assesses whether the current input falls within evaluated conditions. Generalised out-of-distribution detection research indicates that distributional uncertainty includes multiple settings with different assumptions and evaluation requirements. The framework therefore treats out-of-distribution assessment as a separate responsibility rather than a substitute for calibration or predictive uncertainty estimation [26].

The third stage connects confidence calibration, miscalibration review, selective prediction, ambiguity disclosure, and risk communication with lifecycle monitoring. Because model performance may deteriorate as populations, measurements, environments, or relationships change, prediction boundaries cannot remain static. Drift and degradation monitoring should feed back into scope statements, uncertainty assessments, calibration review, and decisions about whether system outputs should continue to be issued [27].

The fourth stage designs the human pathway. Abstention or rejection should be accompanied, where feasible, by an intelligible reason that identifies insufficient evidence, ambiguity, unfamiliarity, or another relevant limitation [28]. The framework also preserves meaningful disagreement rather than automatically reducing contested labels to noise, because disagreement may represent ambiguity, differing expertise, or alternative interpretations [29]. Explanation formats and uncertainty communication should then be evaluated under differing uncertainty conditions and user contexts [30].

Evaluation and Implementation Implications

Evaluation should begin by determining whether identified uncertainty types correspond to the claims made about them. Aleatoric, epistemic, predictive, and distributional uncertainty may require different validation strategies, and their estimates may interact. Confidence calibration should be examined globally and locally, across subgroups, boundary regions, and changing conditions. Miscalibration monitoring should focus particularly on high-confidence errors because these cases may be most likely to promote inappropriate reliance [13].

Selective prediction evaluation should examine both the retained predictions and the cases rejected, deferred, or withheld. Relevant questions include whether errors remain concentrated among retained cases, whether rejection affects populations unequally, and whether escalation pathways provide timely access to better evidence or expertise. Abstention quality cannot be assessed solely by showing reduced automated coverage; downstream review, delay, unresolved cases, and final consequences also require evaluation [14].

Human-centered assessment should evaluate whether ambiguity disclosure, risk communication, and uncertainty visualization support the intended reasoning task without imposing disproportionate cognitive burden. Human-review quality should be assessed through reviewer reasoning, overrides, timeliness, expertise, and outcomes rather than the mere presence of oversight. Trust-calibration and automation-bias evaluation should include both correct and incorrect system advice so that overreliance and underreliance can be distinguished [21].

Distribution-shift monitoring, documentation, auditability, and feedback updating are implementation implications rather than proof of operational readiness. Foundation-model and generative-AI applications may require additional research on open-ended output uncertainty, unsupported responses, explanation instability, and qualified refusal; the present framework supports such investigation but does not validate a universal communication method. These implications are

conceptual research directions, not legal, compliance, procurement, audit, clinical, financial, model-selection, threshold-selection, platform-selection, safety-certification, or deployment instructions.

Conclusion

This article proposed an uncertainty-aware explainable decision-intelligence framework integrating uncertainty-source identification, confidence calibration, ambiguity disclosure, risk communication, selective prediction, abstention, deferral, human review, prediction boundaries, monitoring, documentation, and feedback updating. Its central premise is that trustworthy decision intelligence should communicate what a model does not know rather than presenting a prediction, confidence score, or explanation as sufficient evidence for action. Aleatoric, epistemic, and distributional uncertainty require distinct interpretation, while calibrated confidence, qualified responses, selective prediction, and human escalation require context-specific evaluation. The framework can help structure uncertainty-aware system development and research, but it does not independently establish correctness, safety, decision validity, appropriate reliance, ethical sufficiency, legal compliance, or deployment readiness.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 05 Sep 2025

Revised: 29 Oct 2025

Accepted: 22 Nov 2025

Published online: 15 March 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

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>.

Miller T. Explanation in artificial intelligence: insights from the social sciences. *Artif Intell.* 2019;267:1-38.
<https://doi.org/10.1016/j.artint.2018.07.007>.

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>.

Glikson E, Woolley AW. Human trust in artificial intelligence: review of empirical research. *Acad Manag Ann.* 2020;14(2):627-60.
<https://doi.org/10.5465/annals.2018.0057>.

Thiebes S, Lins S, Sunyaev A. Trustworthy artificial intelligence. *Electron Mark.* 2021;31(2):447-64.
<https://doi.org/10.1007/s12525-020-00441-4>.

Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, et al. A review of uncertainty quantification in deep learning: techniques, applications and challenges. *Inf Fusion.* 2021;76:243-97.
<https://doi.org/10.1016/j.inffus.2021.05.008>.

Mena J, Pujol O, Vitrià J. A survey on uncertainty estimation in deep learning classification systems from a Bayesian perspective. *ACM Comput Surv.* 2022;54(9):1-35.
<https://doi.org/10.1145/3477140>.

Gawlikowski J, Tassi CRN, Ali M, Lee J, Humt M, Feng J, et al. A survey of uncertainty in deep neural networks. *Artif Intell Rev.* 2023;56(Suppl 1):1513-89.
<https://doi.org/10.1007/s10462-023-10562-9>.

Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach Learn.* 2021;110(3):457-506.
<https://doi.org/10.1007/s10994-021-05946-3>.

Hirschfeld L, Swanson K, Yang K, Barzilay R, Coley CW. Uncertainty quantification using neural networks for molecular property prediction. *J Chem Inf Model.* 2020;60(8):3770-80.
<https://doi.org/10.1021/acs.jcim.0c00502>.

Scalia G, Grambow CA, Pernici B, Li YP, Green WH. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. *J Chem Inf Model.* 2020;60(6):2697-717.
<https://doi.org/10.1021/acs.jcim.9b00975>.

Ayhan MS, Kühlewein L, Aliyeva G, Inhoffen W, Ziemssen F, Berens P. Expert-validated estimation of diagnostic uncertainty for deep neural networks in diabetic retinopathy detection. *Med Image Anal.* 2020;64:101724.
<https://doi.org/10.1016/j.media.2020.101724>.

Silva Filho T, Song H, Perello-Nieto M, Santos-Rodriguez R, Kull M, Flach P. Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Mach Learn.* 2023;112(9):3211-60.
<https://doi.org/10.1007/s10994-023-06336-7>.

Hendrickx K, Perini L, Van der Plas D, Meert W, Davis J. Machine learning with a reject option: a survey. *Mach Learn.* 2024;113(5):3073-110.
<https://doi.org/10.1007/s10994-024-06534-x>.

Lahoti P, Gummadi KP, Weikum G. Responsible model deployment via model-agnostic uncertainty learning. *Mach Learn.* 2023;112(3):939-70.
<https://doi.org/10.1007/s10994-022-06248-y>.

Xia G, Bouganis CS. Augmenting the softmax with additional confidence scores for improved selective classification with out-of-distribution data. *Int J Comput Vis*. 2024;132(9):3714-52.
<https://doi.org/10.1007/s11263-024-02029-3>.

Hullman J, Qiao X, Correll M, Kale A, Kay M. In pursuit of error: a survey of uncertainty visualization evaluation. *IEEE Trans Vis Comput Graph*. 2019;25(1):903-13.
<https://doi.org/10.1109/TVCG.2018.2864889>.

Cao S, Liu A, Huang CM. Designing for appropriate reliance: the roles of AI uncertainty presentation, initial user decision, and user demographics in AI-assisted decision-making. *Proc ACM Hum Comput Interact*. 2024;8(CSCW1):1-32.
<https://doi.org/10.1145/3637318>.

Schäfer U, Sipos L, Müller-Birn C. The AI is uncertain, so am I. What now?: navigating shortcomings of uncertainty representations in human-AI collaboration with capability-focused guidance. *Proc ACM Hum Comput Interact*. 2025;9(7):1-48.
<https://doi.org/10.1145/3757451>.

Kale A, Kay M, Hullman J. Visual reasoning strategies for effect size judgments and decisions. *IEEE Trans Vis Comput Graph*. 2021;27(2):272-82.
<https://doi.org/10.1109/TVCG.2020.3030335>.

Green B, Chen Y. The principles and limits of algorithm-in-the-loop decision making. *Proc ACM Hum Comput Interact*. 2019;3(CSCW):1-24.
<https://doi.org/10.1145/3359152>.

Buçinca Z, Malaya MB, Gajos KZ. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proc ACM Hum Comput Interact*. 2021;5(CSCW1):1-21.
<https://doi.org/10.1145/3449287>.

Cao S, Huang CM. Understanding user reliance on AI in assisted decision-making. *Proc ACM Hum Comput Interact*. 2022;6(CSCW2):1-23.
<https://doi.org/10.1145/3555572>.

Chen V, Liao QV, Vaughan JW, Bansal G. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *Proc ACM Hum Comput Interact*. 2023;7(CSCW2):1-32.
<https://doi.org/10.1145/3610219>.

Ruff L, Kauffmann JR, Vandermeulen RA, Montavon G, Samek W, Kloft M, et al. A unifying review of deep and shallow anomaly detection. *Proc IEEE*. 2021;109(5):756-95.
<https://doi.org/10.1109/JPROC.2021.3052449>.

Yang J, Zhou K, Li Y, Liu Z. Generalized out-of-distribution detection: a survey. *Int J Comput Vis*. 2024;132(12):5635-62.
<https://doi.org/10.1007/s11263-024-02117-4>.

Bayram F, Ahmed BS, Kassler A. From concept drift to model degradation: an overview on performance-aware drift detectors. *Knowl Based Syst*. 2022;245:108632.
<https://doi.org/10.1016/j.knosys.2022.108632>.

Artelt A, Visser R, Hammer B. I do not know! but why?—local model-agnostic example-based explanations of reject. *Neurocomputing*. 2023;558:126722.
<https://doi.org/10.1016/j.neucom.2023.126722>.

Uma AN, Fornaciari T, Hovy D, Paun S, Plank B, Poesio M. Learning from disagreement: a survey. *J Artif Intell Res*. 2021;72:1385-470.
<https://doi.org/10.1613/jair.1.12752>.

Cau FM, Hauptmann H, Spano LD, Tintarev N. Effects of AI and logic-style explanations on users' decisions under different levels of uncertainty. *ACM Trans Interact Intell Syst*. 2023;13(4):1-42.
<https://doi.org/10.1145/3588320>.