ORIGINAL RESEARCH **Open access**

Boundary Conditions for Trustworthy Uncertainty Quantification in Materials AI: A Framework for Decision-Ready Confidence

Anna Kowalska[1*], Piotr Nowak[1], Tomasz Zielinski[2], Katarzyna Mazur[1]

Abstract

Uncertainty quantification (UQ) has become a routine component of materials artificial intelligence (AI), with predictive models now systematically reporting confidence intervals or variance estimates alongside outputs spanning formation energies to mechanical properties. Despite this integration, the designation “trustworthy uncertainty” remains conceptually unresolved. Assertions of reliability are frequently decoupled from operational criteria that connect statistical behavior to the concrete decisions faced by materials scientists, including large-scale screening, experimental prioritization, design optimization, certification under regulatory constraints, and the interpretation of anomalous phenomena. Addressing this gap, the present boundary-focused analysis advances a definition of trustworthiness grounded in decision relevance and introduces the construct of decision-ready confidence. This formulation identifies a set of boundary conditions that collectively determine whether uncertainty estimates can support action. Calibration ensures correspondence between predicted uncertainty and empirical error distributions, while sharpness constrains interval width to maintain discriminative value without sacrificing validity. A related requirement concerns the separation of epistemic and aleatoric components, enabling differentiation between reducible and irreducible uncertainty. Coverage, particularly in its conditional form, establishes reliability at the level of individual predictions, and stability enforces robustness under small perturbations of input space. These statistical conditions are complemented by computational tractability, which situates uncertainty estimation within the temporal constraints of decision-making processes. Crucially, none of these properties is intrinsic in isolation; each must be interpreted relative to the decision context in which the model is deployed. To anchor these criteria, the analysis delineates a set of recurring decision regimes that structure materials AI workflows, spanning screening, experimental validation, active learning, optimization, certification, and discovery. Each regime imposes distinct requirements on uncertainty behavior, thereby redefining trustworthiness as a context-dependent alignment rather than a universal attribute. Building on this premise, a framework for decision-ready confidence is introduced to formalize the mapping between decision type, required UQ properties, validation procedures, risk thresholds, and reporting practices. This framework integrates and extends established approaches, including epistemic–aleatoric decomposition, conformal prediction, and broader trustworthy machine-learning paradigms, while situating them within the operational realities of materials engineering. In doing so, it establishes a coherent conceptual foundation for evaluating and deploying UQ methods whose outputs are not only statistically sound but also directly actionable in advancing materials discovery and ensuring system-level reliability.

Keywords Materials artificial intelligence, Trustworthy uncertainty quantification, Decision-ready confidence, Boundary conditions for UQ, Calibration and sharpness, Epistemic-aleatoric separation

*Correspondence:
Anna Kowalska
anna.kowalska@gmail.com
[1] Department of Data-Driven Materials Science, Faculty of Engineering, University of Warsaw, Warsaw, Poland
[2] Department of Computational Materials Engineering, Faculty of Technology, Warsaw University of Technology, Warsaw, Poland

Introduction

Materials AI models now routinely provide uncertainty estimates alongside predictions. Papers claim their UQ is "trustworthy" or "reliable" [1, 2]. But what does that actually mean? For a materials scientist deciding whether to synthesize a candidate, a 0.1 eV/atom uncertainty may be trustworthy. For a nuclear reactor designer, 0.1 eV/atom may be unacceptable. Trustworthiness is not a property of the UQ method alone—it depends on the decision context [3, 4]. This paper provides a boundary/definitional analysis of trustworthy uncertainty for materials AI and proposes a framework for decision-ready confidence.

The past decade has seen an explosion of machine-learning models for materials property prediction, from formation energies and band gaps to elastic moduli and defect migration barriers. Alongside point predictions, researchers have increasingly adopted techniques that also output uncertainty, ranging from Gaussian process regression to deep ensembles, Bayesian neural networks, and conformal prediction wrappers [5-8]. Thuraisingham highlighted the growing emphasis on trustworthy machine learning specifically tailored to materials science challenges [1]. Yet the literature reveals a persistent gap: the same statistical tools that produce excellent calibration on benchmark test sets are often deployed without explicit linkage to the downstream decision they are meant to support [9-11]. Kendall and Gal distinguished epistemic from aleatoric uncertainty in the broader machine-learning community [12], while Deringer and co-workers demonstrated the power of Gaussian processes for materials and molecules [5], but neither work supplied operational criteria for when the resulting uncertainty becomes decision-ready [1, 13].

This ambiguity matters because materials decisions span a wide risk spectrum. In high-throughput virtual screening, an uncertainty estimate only needs to preserve ranking fidelity [14, 15]. In safety-critical certification for nuclear or aerospace applications, the same estimate must carry a conservative coverage guarantee that regulators can audit [3, 16, 17]. A method that performs well on one task can fail catastrophically on another, yet current papers frequently use "trustworthy" as a catch-all adjective without defining the decision boundary [2, 18]. The present work therefore treats "trustworthy uncertainty" as a boundary concept: it exists only when a set of clearly articulated conditions are met relative to a specified decision context [19, 20].

The analysis proceeds in four further steps. Section 2 surveys five distinct ways the term "trustworthy uncertainty" is currently used in the literature and exposes the conceptual confusion each usage creates [6, 12, 21]. Section 3 defines six boundary conditions that any UQ method must satisfy before its outputs can be trusted [9, 13]. Section 4 classifies the major decision types encountered in materials AI and maps each to its minimal UQ requirements [1, 3, 22]. Section 5 assembles these elements into a practical framework for achieving decision-ready confidence, complete with validation protocols, threshold-setting rules, and reporting standards [4, 19, 23]. Throughout, the focus remains strictly conceptual: no new datasets, no performance tables, and no simulations are presented. The goal is to supply the field with precise operational definitions and decision-aware boundary conditions that future UQ development and deployment can reference [1, 2, 24].

By the end of the paper, readers will possess a shared vocabulary and a checklist that replaces vague claims of trustworthiness with auditable, context-specific statements of decision readiness. This shift is essential if materials AI is to move from academic demonstration to industrial adoption and regulatory acceptance [3, 16, 17].

Current Usage of "Trustworthy Uncertainty"

The expression "trustworthy uncertainty" is increasingly invoked in materials machine-learning, yet its semantics fragment across multiple, at times incompatible, interpretations [1, 2, 11]. In some accounts, trustworthiness is equated with calibration, such that nominal confidence levels coincide with empirical error frequencies; Gruich and co-workers, for instance, foreground distribution-specific uncertainty quantification to support neural-network predictions of materials properties [2, 13, 25]. This alignment is necessary but insufficient, as intervals may remain so diffuse that they fail to constrain action [9]. A related interpretation conflates trustworthiness with low predictive variance, where methods such as

deep ensembles are framed as providing scalable uncertainty estimates [6]. Under this view, reduced dispersion is taken as evidence of reliability, although it may simply reflect miscalibrated overconfidence, particularly under distributional shift [12, 13]. Elsewhere, trustworthiness is inferred from downstream use, as when uncertainty informs candidate selection for synthesis; yet such pragmatic invocation leaves unresolved whether uncertainty substantively guided the decision or merely accompanied it [14, 22]. Another strand emphasizes the decomposition of epistemic and aleatoric components, following formulations introduced in computer vision and extended to materials systems [5, 7, 12]. While this distinction clarifies whether uncertainty is reducible, it does not ensure that either component is calibrated or sufficiently informative for the task [1, 21]. Conformal prediction introduces a further perspective by guaranteeing nominal coverage, exemplified in recent materials applications [8, 19, 23]. Although marginal guarantees may obscure local failures where conditional coverage deteriorates [9, 24]. The resulting ambiguity is not merely semantic: models may satisfy one statistical criterion while violating others, leaving their decision relevance indeterminate. Absent an operational linkage between statistical properties and decision consequences, claims of trustworthiness remain contingent and difficult to reproduce [2, 10], motivating the need for explicit boundary conditions [1].

Boundary Conditions for Trustworthy UQ

Trustworthy uncertainty quantification in materials AI is therefore best understood through a set of stringent boundary conditions that articulate how statistical properties translate into decision viability [9, 11]. Calibration requires that predicted uncertainty aligns with empirical error distributions, typically assessed through reliability diagrams approaching unit slope and low expected calibration error; without this correspondence, reported confidence lacks interpretive meaning for any decision [2, 13, 26]. This requirement interacts directly with sharpness, since intervals must be sufficiently narrow to discriminate among alternatives while preserving calibration, otherwise even well-calibrated predictions fail to inform action [13, 14]. A further condition concerns the identifiability of epistemic and aleatoric contributions, which can be operationalized through data-scaling behavior; such separation is indispensable when decisions hinge on whether uncertainty can be reduced through additional data or improved modeling [5, 7, 12]. Coverage must also hold at the level of individual inputs rather than only in aggregate, as marginal guarantees do not preclude localized failure in regions critical to high-stakes decisions [3, 19, 23]. Stability imposes continuity on the uncertainty function, ensuring that small perturbations in input do not induce disproportionate changes in estimated confidence, a prerequisite for optimization and active learning workflows [9, 22, 27]. Finally, computational tractability constrains uncertainty estimation to the temporal scale of the decision loop, since excessive latency renders even statistically sound estimates operationally irrelevant [14, 28]. These conditions jointly define the threshold at which uncertainty becomes actionable; violation of any single constraint undermines its reliability within a given context [1, 2, 4].

Decision Types and their UQ Requirements

The adequacy of these conditions becomes clearer when situated within the spectrum of decision regimes encountered in materials AI, each of which privileges different aspects of uncertainty [1, 10]. In high-throughput screening, the central requirement is the preservation of ranking fidelity across large candidate sets, so that uncertainty need only correlate monotonically with error magnitude rather than achieve strict calibration [9, 14, 15]. This emphasis shifts when selecting candidates for experimental validation, where miscalibration directly translates into resource misallocation, rendering tight calibration and high coverage indispensable [2, 26]. In active learning, the acquisition process depends specifically on epistemic uncertainty, making its separation from irreducible noise a defining requirement [6, 7, 12]. Design optimization introduces further constraints, as algorithms navigating complex design spaces require uncertainty estimates that are simultaneously sharp, calibrated, and locally reliable to avoid infeasible or unstable trajectories [5, 22, 27]. Under safety-critical certification, the tolerance for failure narrows dramatically, elevating conditional coverage and auditability to non-negotiable criteria [3, 16-18]. By contrast, in scientific discovery, uncertainty serves a diagnostic role, where elevated estimates on novel configurations signal departures from the training domain and potential avenues for new physics [15, 19, 21]. These distinctions underscore that trustworthiness is not intrinsic to a method in isolation but emerges from its alignment with the epistemic and operational demands of a specific decision context [1, 20]. **Table 1** defines the minimal

boundary-condition structure required for each decision type, showing that trustworthy uncertainty is inherently decision-conditional rather than method-intrinsic.

Table 1. Decision-Conditional Validity Structure: Minimal Boundary Condition Requirements across Materials AI Decision Types

Decision Type	Calibration	Sharpness	Separation	Conditional Coverage	Stability	Computational Tractability	Critical Failure Mode if Violated
High-Throughput Screening	Medium	Low	Optional	Optional	Low	High	Ranking distortion leading to candidate misprioritization
Experimental Validation Selection	High	Medium	Optional	Medium	Medium	Medium	Resource misallocation due to false positives/negatives
Active Learning Query	Medium	Low	**Mandatory**	Optional	Medium	High	Incorrect acquisition due to conflated uncertainty sources
Design Optimization	High	High	Medium	**Mandatory**	High	Medium	Optimization instability and infeasible design trajectories
Safety-Critical Certification	**Mandatory**	Medium	Medium	**Mandatory (≥99%)**	**Mandatory**	Low	Regulatory rejection or unsafe certification
Scientific Discovery	Medium	Low	Medium	Medium	High	Medium	Failure to detect extrapolation or novel physics

Proposed Framework for Decision-Ready Confidence

Decision-ready confidence is formalized through an integrated framework that aligns uncertainty quantification with explicit decision requirements, thereby translating abstract statistical properties into operational criteria [1, 4]. The process begins with precise specification of the decision context, including the relevant decision regime, admissible risk tolerance, computational constraints, and required coverage level, since these parameters delimit what constitutes acceptable uncertainty in practice [3, 10, 20]. This specification directly conditions the selection of UQ methodology, as different decision environments privilege distinct statistical attributes: ranking-oriented settings may rely on ensemble variance, whereas experimental prioritization benefits from conformal guarantees, active learning depends on isolating epistemic uncertainty, optimization contexts favor Bayesian formulations with subsequent calibration, certification requires conservative coverage augmented by safety margins, and discovery-oriented tasks leverage extrapolation sensitivity combined with ensemble behavior [3, 5, 6, 8, 9, 12, 19, 22-24, 27].

Method selection alone remains insufficient without rigorous validation against the previously defined boundary conditions, which ensures that uncertainty estimates retain their intended meaning under the operational regime. Calibration, sharpness, separability, coverage, and stability must therefore be empirically verified using context-appropriate diagnostics, including reliability analysis, interval-width evaluation, data-scaling behavior, coverage testing, and perturbation sensitivity [2, 13, 26]. These validated properties then inform the construction of decision thresholds, where acceptable risk is translated into concrete operational rules, ranging from ordinal ranking criteria in low-risk settings to stringent high-coverage intervals combined with explicit safety factors under high-stakes conditions [4, 14, 20]. Transparency is enforced through a reporting standard that requires full disclosure of the decision context, validation outcomes, applied thresholds, and known limitations, thereby constraining interpretive ambiguity and enabling reproducibility [1, 16, 17].

Conceptually, the framework can be represented as a mapping from decision regimes to boundary-condition requirements, structured along increasing levels of risk and regulatory scrutiny, with each configuration feeding into a central execution process that sequentially enforces context specification, method selection, validation, thresholding, and reporting. Failures at any stage trigger a return to method selection, ensuring iterative refinement until all conditions are satisfied. Through this recursive structure, the framework redefines trustworthiness as a verifiable property of alignment between uncertainty estimates and decision demands, yielding auditable and context-specific statements of decision readiness rather than abstract or purely statistical claims [19, 23, 24].

Figure 1 presents the conceptual architecture that operationalizes trustworthy uncertainty quantification in materials artificial intelligence.

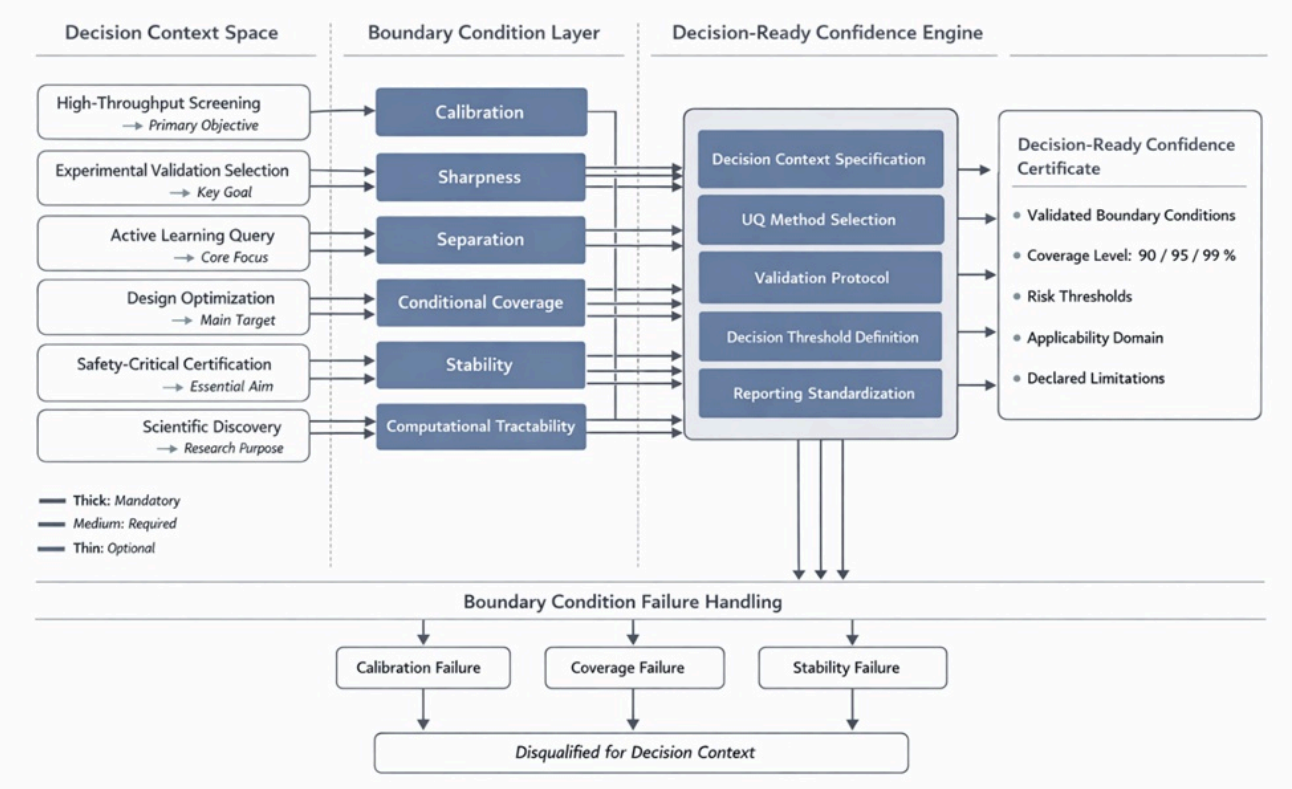

Figure 1. Structural Architecture of Decision-Ready Confidence: Mapping Boundary Conditions to Materials AI Decision Contexts

Boundary Cases and Gray Zones

Even when the six boundary conditions are clearly stated, real-world applications frequently produce gray zones where a UQ method satisfies some conditions but not others. These boundary cases illustrate why trustworthiness must always be evaluated relative to a specific decision context rather than in absolute terms [1, 9].

Table 2 formalizes the failure modes of UQ by linking each violated boundary condition to its statistical signature and its concrete decision-level consequence.

Table 2. Structural Failure Taxonomy of Uncertainty Quantification: Mapping Boundary Violations to Decision-Level Consequences

Boundary Condition Violated	Statistical Symptom	Hidden Risk	Decision-Level Consequence	Detectability	Recoverability
Calibration Failure	Reliability curve deviates from diagonal	Misleading confidence intervals	All decisions invalidated	High	Moderate (recalibration)
Sharpness Failure	Excessively wide intervals	Loss of discriminative power	Inefficient screening and optimization stagnation	High	Low
Separation Failure	Epistemic does not decay with data	Misidentification of reducible uncertainty	Active learning collapse	Medium	Moderate
Conditional Coverage Failure	Local undercoverage	Hidden high-risk regions	Catastrophic errors in certification	Low	Low
Stability Failure	High sensitivity to small perturbations	Non-robust uncertainty signals	Optimization and decision volatility	Medium	Low
Computational Tractability Failure	High latency	Temporal mismatch with decision loop	Infeasible deployment in real workflows	High	High

Case Analysis: Context-Dependent Validity of UQ

These cases delineate the operational limits within which uncertainty estimates retain decision relevance. Situations in which calibration is achieved but sharpness is deficient illustrate how statistical validity does not guarantee utility: excessively wide intervals preserve ranking fidelity and may remain acceptable in high-throughput screening, yet they undermine optimization and certification by failing to constrain actionable trade-offs [3, 9, 13, 14]. The inverse configuration, where intervals are narrow but miscalibrated, represents a more fundamental breakdown, as overconfidence severs the link between reported uncertainty and empirical error, rendering the estimates unusable across all decision contexts regardless of apparent precision [2, 13]. A related tension emerges in conformal settings where marginal coverage is satisfied while conditional coverage deteriorates in localized regions; such behavior may suffice for aggregate screening but becomes untenable when decisions hinge on individual candidates within specific compositional or structural neighborhoods [8, 19, 23, 24]. Claims of epistemic–aleatoric separation without empirical validation further exemplify how nominal methodological properties fail to translate into decision support when the underlying assumptions remain untested, particularly in active-learning regimes that depend on reducible uncertainty signals [5, 7, 11, 12]. Even when all boundary conditions hold under interpolation, their breakdown under extrapolation exposes the fragility of

trustworthiness outside a defined applicability domain, necessitating explicit detection and re-validation mechanisms [9, 15]. These boundary cases underscore that trustworthiness is inherently conditional, with identical UQ methods satisfying some decision regimes while failing others, thereby reinforcing the primacy of context specification prior to deployment [1, 6, 10].

Relation to Existing Frameworks

The proposed framework extends established uncertainty quantification paradigms by embedding them within a decision-aware structure tailored to materials AI. Conformal prediction contributes finite-sample coverage guarantees that have demonstrated utility in materials property prediction [8, 19, 23], yet its emphasis on marginal validity leaves unresolved questions of sharpness, separability, and robustness under perturbation [24]. Within the present formulation, conformal methods are retained as critical components for experimental selection and certification, but their outputs are interpreted through additional boundary conditions that ensure decision relevance [1, 9, 16]. The epistemic–aleatoric distinction, originating in computer vision and subsequently adapted to materials modeling through Gaussian-process approaches [5, 7, 12], provides a principled mechanism for distinguishing reducible from irreducible uncertainty. This distinction is preserved but operationalized through explicit validation and alignment with decision types, ensuring that epistemic uncertainty informs acquisition strategies while aleatoric components guide risk assessment [11, 21]. Extrapolation detection frameworks further contribute by identifying departures from the training domain, a capability integrated here into both discovery-oriented decisions and stability requirements [9, 15]. Broader trustworthy-machine-learning approaches, including deep ensembles and Bayesian uncertainty taxonomies [4, 6, 12], offer mature statistical tools, yet their original development contexts differ substantially from materials applications. By mapping these methods onto materials-specific decision regimes and boundary conditions, the framework reconciles general-purpose uncertainty modeling with the practical demands of synthesis, optimization, and certification [1, 3, 22, 26]. In this sense, it functions not as a replacement but as an interpretive layer that renders existing methods operationally meaningful.

Implications for UQ Method Development

The framework reorients uncertainty quantification research toward explicit decision alignment, introducing concrete expectations for methodological development and evaluation. Any proposed method must articulate its intended decision context at the outset, as abstract claims of trustworthiness without such specification preclude meaningful assessment of applicability [1, 2, 4, 10]. This shift extends to validation practices, which must systematically address all relevant boundary conditions rather than selectively emphasizing isolated metrics such as calibration or coverage; comprehensive evaluation now entails reliability analysis, interval-width characterization, empirical tests of uncertainty decomposition, localized coverage assessment, stability under perturbation, and computational feasibility within the target decision loop [7, 9, 13, 19, 26, 27]. Reporting standards are similarly elevated, requiring transparent disclosure of context, validation outcomes, operational thresholds, and explicit limitations, thereby constraining overgeneralization and enabling reproducibility across application domains [3, 16-18]. While these requirements increase the evidentiary burden, they also enhance the immediate usability of UQ methods, allowing practitioners to deploy them with confidence in contexts where boundary conditions are demonstrably satisfied [14, 15, 20, 22]. Editorial and review processes can leverage this structure to transform assessments of trustworthiness from subjective judgment into verifiable criteria grounded in decision relevance [1, 21]. The broader implication is a shift from methodological abstraction toward integrated decision support, accelerating the translation of materials AI into practice.

Recommendations for Practice

Adoption of the framework enables coordinated improvements across methodological development, benchmarking, and regulatory evaluation by aligning uncertainty quantification with decision-specific requirements. Method developers are encouraged to design approaches with explicit target contexts and to provide verifiable calibration guarantees across

defined risk levels, accompanied by modular validation procedures that facilitate downstream applicability checks [8, 9, 13, 19, 23]. Benchmark design similarly benefits from a transition toward decision-aware evaluation, replacing aggregate error metrics with performance criteria tied to distinct operational regimes, including ranking fidelity, resource allocation accuracy, data-efficiency gains, constraint satisfaction, certification-grade coverage, and extrapolation sensitivity [2, 14, 15, 22, 26]. Regulatory applications, particularly in safety-critical domains, can operationalize the framework by specifying auditable requirements for coverage, stability, and transparency, thereby replacing ambiguous notions of trustworthy AI with enforceable standards grounded in materials-specific risk considerations [3, 16-18]. At the level of research practice, early-stage specification of decision context prior to model development ensures methodological alignment and mitigates misapplication across incompatible regimes [4, 10, 20]. Through these coordinated shifts, uncertainty quantification becomes a structured component of decision-making infrastructure, supporting the emergence of materials AI systems whose outputs meet the reliability expectations traditionally associated with established engineering and experimental methodologies [28, 29].

Conclusion

“Trustworthy uncertainty” has persisted as an aspirational yet under-specified construct in materials AI, limiting its interpretability and practical deployment. The present analysis resolves this ambiguity by introducing a set of boundary conditions—calibration, sharpness, separation, conditional coverage, stability, and computational tractability—together with a corresponding typology of decision regimes spanning screening, experimental selection, active learning, optimization, certification, and discovery. Framed in this way, uncertainty acquires meaning only through its alignment with the statistical and operational demands of a given decision context.

Building on this foundation, the decision-ready confidence framework formalizes how uncertainty estimates are rendered actionable. Explicit specification of decision context constrains method selection, while systematic validation against relevant boundary conditions ensures that statistical properties translate into operational reliability. This process is completed through the imposition of risk-calibrated thresholds and transparent reporting standards that expose both capabilities and limitations. The outcome is a shift from abstract claims toward auditable statements of decision readiness, grounded in measurable criteria rather than interpretive judgment.

This perspective also clarifies the relationship between general-purpose trustworthy-machine-learning methods and the domain-specific requirements of materials science. By embedding established techniques within a decision-aware structure, the framework reconciles statistical rigor with the constraints of synthesis cost, regulatory oversight, multi-objective optimization, and exploratory discovery. Trustworthiness thus emerges not as an intrinsic attribute of an algorithm but as a relational property defined by the interaction between method, context, and acceptable risk.

Widespread adoption of this framework would standardize both terminology and evaluation, enabling cumulative progress across the field. When uncertainty quantification studies consistently declare their target decision regimes, validate the corresponding boundary conditions, and report limitations with precision, materials AI can transition from proof-of-concept demonstrations to dependable decision-support systems. The conceptual infrastructure is now established; its impact depends on disciplined and collective implementation.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 30 Apr 2023 Revised: 13 Aug 2023 Accepted: 26 Nov 2023
Published online: 18 January 2024

References

Thuraisingham B. Trustworthy machine learning. IEEE Intell Syst. 2022;37(1):21-4.
https://doi.org/10.1109/MIS.2022.3152946.

Gruich CJ, Madhavan V, Wang Y, Goldsmith BR. Clarifying trust of materials property predictions using neural networks with distribution-specific uncertainty quantification. Mach Learn Sci Technol. 2023;4(2):025019.
https://doi.org/10.1088/2632-2153/accace.

Neudecker D, Grosskopf M, Herman M, Haeck W, Grechanuk P, Vander Wiel SA, et al. Enhancing nuclear data validation analysis by using machine learning. Nucl Data Sheets. 2020;167:36-60.
https://doi.org/10.1016/j.nds.2020.07.002.

Zhang X, Chan FT, Yan C, Bose I. Towards risk-aware artificial intelligence and machine learning systems: An overview. Decis Support Syst. 2022;159:113800.
https://doi.org/10.1016/j.dss.2022.113800.

Deringer VL, Bartók AP, Bernstein N, Wilkins DM, Ceriotti M, Csányi G. Gaussian process regression for materials and molecules. Chem Rev. 2021;121(16):10073-141.
https://doi.org/10.1021/acs.chemrev.1c00022.

Lakshminarayanan B, Pritzel A, Blundell C. Simple and scalable predictive uncertainty estimation using deep ensembles. Adv Neural Inf Process Syst. 2017;30.
https://doi.org/10.5555/3295222.3295387.

Li L, Chang J, Vakanski A, Wang Y, Yao T, Xian M. Uncertainty quantification in multivariable regression for material property prediction with Bayesian neural networks. Sci Rep. 2024;14(1):10543.
https://doi.org/10.1038/s41598-024-61189-x.

Angelopoulos AN, Bates S. Conformal prediction: A gentle introduction. Found Trends Mach Learn. 2023;16(4):494-591.
https://doi.org/10.1561/2200000101.

Tran K, Neiswanger W, Yoon J, Zhang Q, Xing E, Ulissi ZW. Methods for comparing uncertainty quantifications for material property predictions. Mach Learn Sci Technol. 2020;1(2):025006.
https://doi.org/10.1088/2632-2153/ab7e1a.

Wu J, Shang S. Managing uncertainty in AI-enabled decision making and achieving sustainability. Sustainability (Basel). 2020;12(21):8758.
https://doi.org/10.3390/su12218758.

Kläs M, Vollmer AM. Uncertainty in machine learning applications: A practice-driven classification of uncertainty. In: Computer safety, reliability, and security. SAFECOMP 2018 Workshops, ASSURE, DECSoS, SASSUR, STRIVE, and WAISE, Västerås, Sweden, September 18, 2018, Proceedings. Cham: Springer; 2018. p. 431-8.
https://doi.org/10.1007/978-3-319-99229-7_36.

Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision? Adv Neural Inf Process Syst. 2017;30.
https://doi.org/10.5555/3295222.3295309.

Pernot P. Calibration in machine learning uncertainty quantification: Beyond consistency to target adaptivity. APL Mach Learn. 2023;1(4):046121.
https://doi.org/10.1063/5.0174943.

Tavazza F, DeCost B, Choudhary K. Uncertainty prediction for machine learning models of material properties. ACS Omega. 2021;6(48):32431-40.
https://doi.org/10.1021/acsomega.1c03752.

Mulukutla M, Robinson R, Khatamsaz D, Vela B, Vu N, Arróyave R. Supply risk-aware alloy discovery and design. arXiv [Preprint]. 2024.
https://doi.org/10.48550/arXiv.2409.15391.

Tambon F, Laberge G, An L, Nikanjam A, Mindom PS, Pequignot Y, et al. How to certify machine learning based safety-critical systems? A systematic literature review. Autom Softw Eng. 2022;29(2):38.
https://doi.org/10.1007/s10515-022-00337-x.

Damiani E, Ardagna CA. Certified machine-learning models. In: SOFSEM 2020: Theory and Practice of Computer Science. Cham: Springer; 2020. p. 3-15.
https://doi.org/10.1007/978-3-030-38919-2_1.

Najar M, Wang H. Establishing operator trust in machine learning for enhanced reliability and safety in nuclear power plants. Prog Nucl Energy. 2024;173:105280.
https://doi.org/10.1016/j.pnucene.2024.105280.

Fontana M, Zeni G, Vantini S. Conformal prediction: A unified review of theory and new challenges. Bernoulli. 2023;29(1):1-23.
https://doi.org/10.3150/21-BEJ1447.

Ahmadi M, Rosolia U, Ingham MD, Murray RM, Ames AD. Risk-averse decision making under uncertainty. IEEE Trans Autom Control. 2024;69(1):55-68.
https://doi.org/10.1109/TAC.2023.3264178.

Kaplan L, Cerutti F, Sensoy M, Preece A, Sullivan P. Uncertainty aware AI ML: Why and how. arXiv [Preprint]. 2018.
https://doi.org/10.48550/arXiv.1809.07882.

Zuo Y, Qin M, Chen C, Ye W, Li X, Luo J, et al. Accelerating materials discovery with Bayesian optimization and graph deep learning. Mater Today. 2021;51:126-35.
https://doi.org/10.1016/j.mattod.2021.08.012.

Angelopoulos AN, Barber RF, Bates S. Theoretical foundations of conformal prediction. arXiv [Preprint]. 2024. https://doi.org/10.48550/arXiv.2411.11824.

Kiyani S, Pappas G, Hassani H. Conformal prediction with learned features. arXiv [Preprint]. 2024. https://doi.org/10.48550/arXiv.2404.17487.

Gurney K. An introduction to neural networks. London: CRC Press; 2018. https://doi.org/10.1201/9781315273570.

Pitz E, Pochiraju K. AI/ML for quantification and calibration of property uncertainty in composites. In: Machine Learning Applied to Composite Materials. Singapore: Springer Nature Singapore; 2022. p. 45-76. https://doi.org/10.1007/978-981-19-6278-3_3.

Manfredi P. Conservative Gaussian process models for uncertainty quantification and Bayesian optimization in signal integrity applications. IEEE Trans Compon Packag Manuf Technol. 2024;14(7):1261-72. https://doi.org/10.1109/TCPMT.2024.3390402.

Islam MM, Moury RK, Pinky KN. Machine learning–driven forecasting pipelines for financial volatility detection in integrated enterprise ERP environments. AJATES. 2022;2(02):134-73. https://doi.org/10.63125/y42nk811.

Pal T, Islam MM, Amin A. Artificial intelligence-based decision support systems and managerial performance. AJATES. 2024;4(04):154-84. https://doi.org/10.63125/wmz8dc67.