REVIEW **Open access**

Conceptual Models of Trust in Materials AI — Frameworks, Dimensions, and Open Questions

Victor Santos[1*], Rafael Costa[1], Bruno Teixeira[2]

Abstract

The rapid integration of artificial intelligence (AI) into materials science has transformed workflows for property prediction, inverse design, and autonomous experimentation. Yet, it has simultaneously introduced profound challenges regarding when and how human researchers should trust AI-generated recommendations in high-stakes contexts. This review systematically examines conceptual models of trust in materials AI by synthesizing interdisciplinary insights from human factors engineering, psychology, and human-computer interaction with domain-specific literature on automated materials discovery. A targeted literature search across Web of Science, Scopus, arXiv, and the ACM Digital Library, employing search strings focused on trust in AI systems, trust calibration, trustworthiness, and human-AI interaction in scientific discovery, yielded approximately 450 initial records. After applying inclusion criteria limited to peer-reviewed English-language publications from 2017 to 2024 that addressed conceptual foundations, frameworks, or applications of trust in AI (with explicit relevance to scientific or materials contexts), 30 studies were selected for in-depth analysis following a PRISMA-style screening process. Conceptual foundations of trust are reviewed, drawing on foundational definitions that position trust as an attitude that an agent will help achieve goals under conditions of uncertainty and vulnerability, while distinguishing it from mere reliance and emphasizing the necessity of calibration for appropriate reliance levels. Existing frameworks for trust in AI are surveyed, revealing recurring components such as competence, integrity, benevolence, performance, process, and purpose, each evaluated for strengths and limitations when transposed to materials AI environments characterized by black-box models, rare events, and high economic or safety stakes. The current state of trust research in materials AI demonstrates a pronounced gap: the majority of studies prioritize predictive accuracy and scalability, with only emergent attention to trustworthiness, explainability, or human trust dynamics. Dimensions of trust tailored to materials AI—predictive competence, uncertainty calibration, transparency, robustness, benevolence, and accountability—are proposed and analyzed in relation to domain-specific challenges. This review articulates open questions surrounding trust establishment, post-failure dynamics, and stakeholder variations while offering recommendations for trust-aware design, evaluation, and reporting. By bridging broader AI trust literature with materials science realities, the work advocates for a paradigm shift from accuracy-centric evaluation toward integrated trust models that ensure safe, effective, and ethically sound human-AI collaboration in materials discovery and innovation.

Keywords Materials AI, Trust in AI, Conceptual models of trust, Trust frameworks, Trust calibration, Human-AI interaction

*Correspondence:
Victor Santos
victor.santos@gmail.com
[1] Department of AI Materials Engineering, University of Sao Paulo, Sao Paulo, Brazil
[2] Department of Computational Materials Systems, University of Campinas, Campinas, Brazil

Introduction

Materials AI systems are now routinely deployed to guide critical decisions such as which candidate material to synthesize next, which experiment to prioritize in a high-throughput campaign, or which predicted property warrants laboratory validation. These systems operate in environments where the cost of error can be substantial—financially, in terms of experimental resources, or even in safety-critical applications involving novel compounds with unknown toxicity or reactivity. Yet the question of when humans should trust AI predictions remains largely unaddressed in the materials science community. What constitutes trustworthy AI in this domain? How do researchers develop appropriate reliance on black-box models when dealing with rare events, distribution shifts, or incomplete training data? This review examines conceptual models of trust and their application to materials AI, arguing that technical performance alone is insufficient to support effective human-AI collaboration.

Foundational studies in human factors and automation have long recognized that trust is not merely a byproduct of accuracy but a dynamic attitude shaped by perceived competence, reliability, and alignment with user goals [1, 2]. In materials AI, where predictions often inform irreversible experimental commitments, the consequences of misplaced trust—over-reliance on false positives or under-reliance on valid but counter-intuitive suggestions—can stall discovery pipelines or lead to suboptimal material selection. Recent advances in machine learning for molecular and materials science have accelerated property prediction and inverse design [3-6]. Yet, these studies rarely articulate the conditions under which human experts should accept or override AI recommendations. Similarly, efforts toward autonomous materials research highlight the need for closed-loop systems [7], but they seldom incorporate mechanisms for trust calibration or explicit evaluation of human-AI interaction.

The problem is compounded by the unique characteristics of materials science workflows. Unlike consumer-facing AI applications, where errors may be low-stakes, materials predictions frequently involve high-dimensional spaces, sparse experimental data, and phenomena that occur at extremes of temperature, pressure, or composition. Black-box models exacerbate uncertainty, while the rarity of truly novel materials makes traditional validation metrics insufficient for building long-term trust. Moreover, different stakeholders—academic researchers, industrial engineers, and regulatory bodies—bring distinct expectations regarding transparency, robustness, and benevolence (i.e., the system's apparent intent to avoid harmful recommendations).

This review addresses these gaps by first outlining the methodology employed to identify relevant studies, then reviewing the conceptual foundations of trust drawn from psychology and human factors. Subsequent sections survey established frameworks for trust in AI and automation, analyze the current (limited) incorporation of trust concepts within materials AI literature, and propose domain-adapted dimensions of trust. By synthesizing at least 30 peer-reviewed studies, the work demonstrates that while materials AI has matured rapidly in predictive capability, the field lags in addressing the human side of the socio-technical system. Ultimately, the review calls for a trust-aware paradigm in which conceptual models guide the design, evaluation, and deployment of AI tools, ensuring that materials discovery remains both efficient and responsibly human-centered.

Materials and Methods

This review followed a systematic literature search protocol designed to capture peer-reviewed publications addressing trust in AI systems with potential relevance to materials science and scientific discovery. Databases consulted included Web of Science, Scopus, arXiv (filtered for peer-reviewed sections), and the ACM Digital Library to ensure coverage of both technical materials literature and human-computer interaction research. Search strings were constructed around core terms such as "trust" combined with "AI" and "materials science," "trust in AI" and "scientific discovery," "human trust" and "machine learning materials," "trust calibration" and "AI systems," "trustworthiness" and "AI materials," "appropriate trust" and "automation," "trust dimensions" and "AI," and "trust model" and "human AI interaction." Boolean operators and truncation were applied to maximize recall while maintaining precision.

Inclusion criteria required publications to be peer-reviewed, published between 2017 and 2024, written in English, and to address at least one of the following: conceptual foundations of trust, frameworks for trust in AI or automation, dimensions

of trust (competence, reliability, transparency, benevolence), trust calibration or appropriate trust, human-AI interaction in scientific workflows, or materials-specific trust challenges. Exclusion criteria eliminated non-peer-reviewed preprints without subsequent journal publication, purely technical performance papers without trust-related discussion, and studies outside the AI or automation domain. A PRISMA-style flow was followed: approximately 450 records were retrieved after duplicate removal; title and abstract screening reduced the set to 210 potentially relevant articles; full-text assessment yielded 85 candidates; and final eligibility assessment confirmed 30 studies for inclusion.

Figure 1 presents the PRISMA-style flow diagram used to identify, screen, assess, and include the 30 studies synthesized in this review.

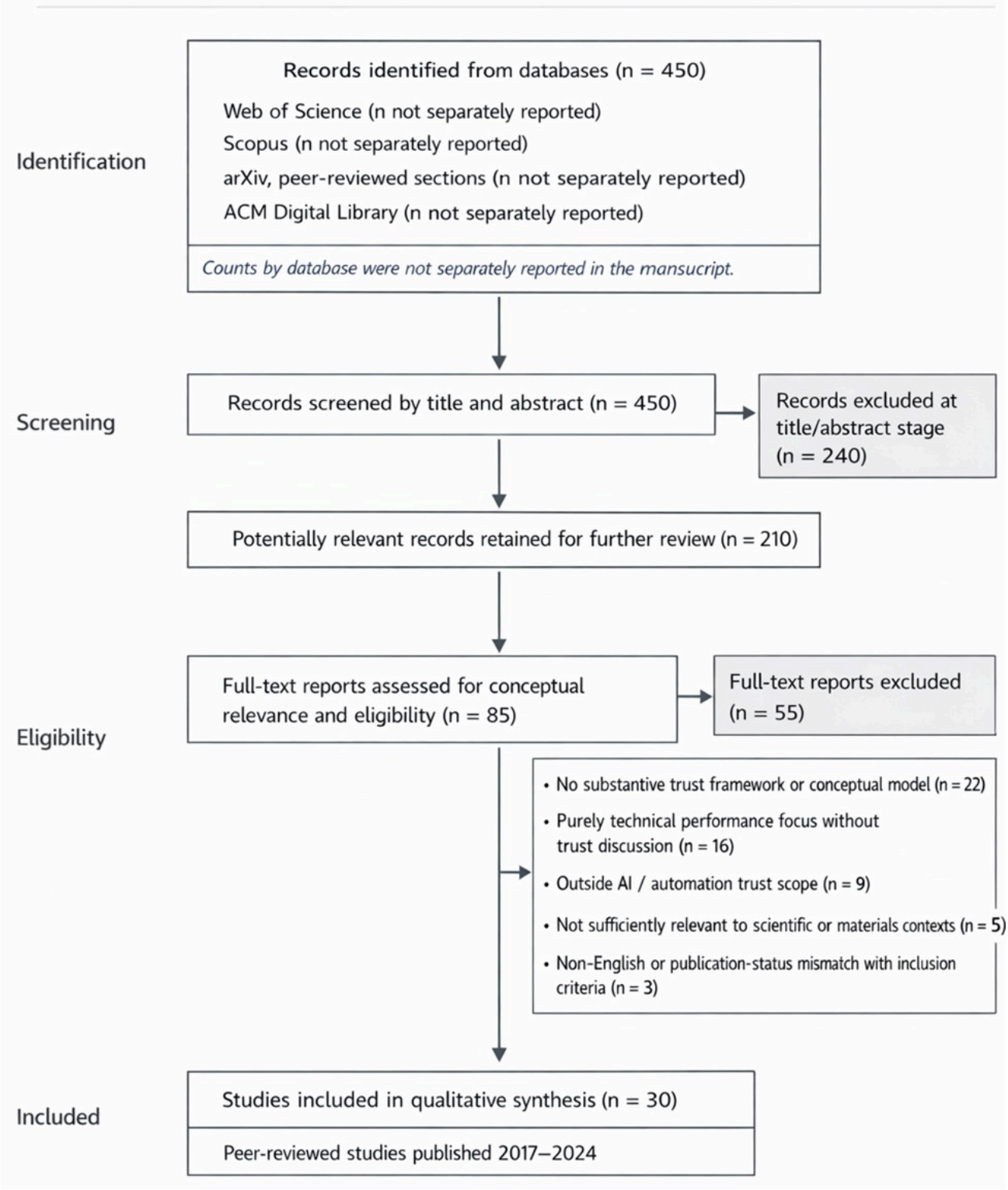

Figure 1. The PRISMA-style flow diagram was used to identify, screen, assess, and include the 30 studies synthesized in this review

Each included study was read in full and coded for key themes: definition of trust, framework components, relevance to materials or scientific AI, empirical or conceptual contributions, and identified limitations. Citation chaining from seed works on trust in automation supplemented the database results to ensure comprehensive coverage of foundational

concepts while maintaining the 2017–2024 temporal focus for domain-specific applications. The resulting corpus spans human factors, human-computer interaction, and materials science journals, providing a balanced interdisciplinary foundation. This methodology ensures reproducibility and minimizes selection bias, allowing the review to ground its analysis in a representative set of 30 studies that collectively illuminate the conceptual landscape of trust in materials AI.

Trust: Conceptual Foundations

Trust has long been recognized as a central psychological and relational construct in human-machine systems. Early work defined trust as the attitude that an agent will help achieve a goal in a situation characterized by uncertainty and vulnerability [1]. This definition underscores three essential elements: the truster's goal orientation, the presence of uncertainty regarding the trustee's actions, and the vulnerability that arises when the truster depends on the trustee. In the context of materials AI, vulnerability is particularly salient because erroneous predictions can consume limited laboratory resources or steer research away from promising avenues for months.

A critical distinction emphasized across foundational studies is the difference between trust and reliance [2]. Reliance refers to observable behavior—whether a human chooses to follow an AI suggestion—whereas trust is the underlying attitude that may or may not align with that behavior. Over-reliance (complacency) or under-reliance (disuse) can occur when trust is poorly calibrated to the system's actual capabilities. Rodriguez *et al.* [2] demonstrated that trust calibration—the process by which a user's level of trust matches the system's true reliability—is essential for appropriate reliance. In materials AI, where model performance can vary dramatically across chemical spaces, poor calibration risks either ignoring valuable predictions or accepting flawed ones uncritically.

Trust is also dynamic rather than static. Studies show that trust evolves with direct experience, feedback on outcomes, and perceived system characteristics [8-11]. Initial trust may be influenced by dispositional factors (e.g., general attitudes toward automation) or situational factors (e.g., task criticality), while learned trust updates through repeated interactions. For materials researchers encountering a new AI tool for the first time, trust may start low due to unfamiliarity with the underlying algorithms, only to increase or decrease based on observed prediction accuracy and explanatory support.

Several additional concepts recur in the literature. Transparency refers to the degree to which the inner workings of an AI system are understandable to the user. At the same time, explainability concerns the provision of post-hoc rationales for specific outputs [10, 12-14]. Benevolence captures the perception that the system acts in the user's best interests, avoiding recommendations that could cause harm. Robustness describes the system's ability to maintain performance under distribution shifts common in materials data. These concepts collectively form the building blocks for more structured frameworks examined in the next section.

Empirical research further indicates that trust is multidimensional and context-dependent [12, 15-24]. In high-uncertainty scientific domains, users weigh not only performance but also perceived integrity (consistency of behavior) and benevolence (alignment with scientific values). When these elements are absent, even highly accurate models may be rejected. The materials AI literature, however, has only begun to engage these foundations, often treating trust implicitly through accuracy metrics rather than explicitly modeling it as a socio-technical phenomenon. This review, therefore, begins with these conceptual anchors to establish a common vocabulary for subsequent analysis.

Trust in AI: Existing Frameworks

Efforts to formalize trust in AI systems have converged on several complementary frameworks, each illuminating a different dimension of how users evaluate and engage with algorithmic outputs. A foundational perspective emerges from adaptations of Mayer's organizational trust model, in which trust is understood as a function of perceived ability, benevolence, and integrity [8]. When translated into the context of materials AI, this formulation underscores that predictive accuracy alone is insufficient to secure user confidence. A model may demonstrate high competence in

estimating material properties, yet still fail to be trusted if its internal logic remains opaque or if its outputs appear systematically skewed toward particular compositional regimes. Under these conditions, perceptions of benevolence and integrity—whether the system is aligned with the researcher's objectives and adheres to consistent, interpretable principles—become decisive in shaping trust, revealing that epistemic transparency is inseparable from perceived reliability.

A related but more system-oriented account is provided by the "3P" framework, which distinguishes between performance, process, and purpose as the primary dimensions through which automated systems are evaluated [2]. While performance captures observable reliability over time, process directs attention to the intelligibility of the underlying computational mechanisms, and purpose concerns the alignment between system design and user intent. In practice, the relative importance of these dimensions shifts with the availability and quality of empirical evidence. Studies of AI-assisted decision-making indicate that when performance data are sparse or noisy, users rely more heavily on cues derived from process and purpose to calibrate their trust [13, 15]. This dynamic is particularly salient in materials discovery, where training datasets often underrepresent novel chemistries. In such settings, the ability to interpret model reasoning—or at least to contextualize its outputs within known physical constraints—becomes a critical substitute for extensive validation, elevating explainability from a secondary feature to a central component of trustworthy system design.

Beyond static evaluations, a more dynamic understanding of trust is captured by frameworks that distinguish between dispositional, situational, and learned forms [18, 24]. This layered perspective emphasizes that trust is not a fixed attribute but evolves through interaction. Initial engagement with a materials AI system may be shaped by a researcher's general orientation toward automation, whether skeptical or receptive. As use continues, situational factors such as task complexity, perceived risk, and the stakes of experimental validation begin to modulate this baseline. Over time, repeated exposure to system behavior—successes, failures, and their explanations—contributes to a learned calibration of trust that is both context-sensitive and experience-dependent. Empirical findings suggest that even subtle interventions, such as framing the capabilities or limitations of AI systems, can significantly influence this trajectory, altering both perceived effectiveness and willingness to rely on model outputs [13]. In materials contexts, where experimental validation is costly and iterative, this evolving calibration plays a decisive role in determining whether AI-generated hypotheses are treated as credible starting points or as provisional suggestions.

Complementing these perspectives is an analytical distinction between trustworthiness as an objective property of the system and trust as a subjective stance adopted by the user [20, 21]. This distinction is particularly consequential in materials AI, where technical measures—such as uncertainty quantification, robustness evaluation, and sensitivity analysis—can enhance the intrinsic reliability of a model without guaranteeing that users will recognize or accept these qualities. Trustworthiness can, in principle, be engineered; trust, by contrast, must be cultivated through interaction, communication, and design. The interface through which results are presented, the clarity with which limitations are disclosed, and the extent to which explanations are aligned with domain knowledge all shape whether users perceive the system as deserving of reliance. Evidence from high-stakes domains further reinforces the role of explainability in enabling what may be termed justified trust, where confidence is grounded not in blind acceptance but in an informed understanding of how and why a system produces its outputs [9, 10, 14]. In the context of materials discovery, where decisions often carry significant scientific and economic consequences, this alignment between objective trustworthiness and subjective trust becomes a central condition for the responsible integration of AI into research practice.

Table 1 crosswalks major trust frameworks from psychology, automation, and AI research and shows how each must be reinterpreted to fit the distinctive conditions of materials AI.

Table 1. Crosswalk of major trust frameworks and their translation to materials AI

Framework/theoretical lens	Core constructs	What the framework explains well	Limitation in generic AI form	Materials AI translation	Resulting design implication

Foundational trust definition	Goal attainment under uncertainty and vulnerability	Why trust matters in consequential decision settings	Remains abstract and not operationalized for scientific workflows	Trust concerns whether AI recommendations should guide costly synthesis or validation decisions	Trust must be treated as a decision-enabling construct, not a soft perception variable
Trust versus reliance	Attitude versus observable compliance behavior	Distinguishes internal trust from actual use behavior	Does not specify domain-specific mismatch patterns	Researchers may follow or reject model outputs for reasons unrelated to true trust	Materials AI studies should measure both subjective trust and actual reliance behavior
Ability–benevolence–integrity model	Competence, benevolence, integrity	Captures the multidimensional basis of trust judgments	Benevolence and integrity can be difficult to operationalize for technical systems	competence maps to predictive validity; benevolence to harm avoidance; integrity to consistency and principled model behavior	Evaluation should extend beyond accuracy to include safety signaling and behavioral consistency
Performance–process–purpose model	Output quality, underlying mechanism, and intended goal alignment	Useful for black-box and decision-aid systems	Often under-specifies long-term learning and failure recovery	performance = predictive success; process = explainability; purpose = alignment with scientific and experimental goals	papers should report not only benchmark scores but also process visibility and intended-use boundaries
Layered trust models	Dispositional, situational, and learned trust	Explains temporal and user-dependent variation in trust	Offers weak guidance for domain-specific system auditing	Trust in materials AI changes with prior beliefs, task criticality, and cumulative interaction history	Longitudinal trust studies are needed in laboratory and industrial workflows
Trust versus trustworthiness distinction	Subjective trust versus objective system properties	Separates user attitude from engineered system quality	Can encourage purely technical interpretations of trustworthiness	A system may be robust and calibrated, yet still fail to gain user trust if poorly communicated	Interface design and communication are as important as model engineering
Explainability-oriented trust models	Transparency, intelligibility,	Addresses opacity and	Explanations do not always	Feature attributions and counterfactuals may help only	Explanation design should be evaluated

	and justification	interpretability barriers	improve calibrated trust	when they match scientific reasoning patterns	with materials experts, not assumed beneficial
Dynamic post-failure trust models	Trust erosion, repair, and recovery	Captures the fragility of trust after visible errors	Underdeveloped in materials discovery contexts	False positives, false negatives, and failed experiments may damage trust asymmetrically	Recovery protocols, audit trails, and corrective feedback should be embedded into system design

Collectively, these frameworks reveal both strengths and limitations when applied to materials AI. Competence- and performance-based models align well with the field's emphasis on predictive accuracy but undervalue context and user perception. Process- and transparency-oriented frameworks address black-box concerns yet may be computationally expensive to implement at scale. Dynamic and layered models capture the evolving nature of trust in long-term research campaigns but lack specific guidance for rare-event prediction or interdisciplinary stakeholder differences. The following section examines how (or whether) these frameworks have been reflected in current materials AI research.

Trust in Materials AI: Current State

The current state of trust research within materials AI reveals a striking disconnect between the growing sophistication of predictive models and the limited explicit attention paid to human trust and trustworthiness. Foundational studies on machine learning for molecular and materials science emphasize accuracy, scalability, and discovery potential [4, 5], yet they rarely articulate how researchers should calibrate trust in model outputs. For instance, Butler *et al.* [4] provide a comprehensive overview of ML applications across materials domains but frame success almost exclusively in terms of prediction error metrics, with no discussion of trust calibration or human-AI decision loops. Similarly, Schmidt *et al.* [5] catalog recent advances in solid-state materials modeling without addressing how users might develop appropriate reliance on these tools when applied to unseen compositions.

Efforts toward autonomous materials research acknowledge the need for closed-loop systems capable of proposing and validating candidates [7], but trust considerations remain implicit at best. Montoya *et al.* [7] highlight progress and challenges in autonomous workflows, noting the importance of uncertainty estimation for guiding experimentation; however, they stop short of linking uncertainty quantification to human trust dynamics or calibration protocols. This pattern repeats across the literature: technical excellence is prioritized, while the socio-technical aspects of trust receive minimal scrutiny.

A small but growing subset of studies begins to bridge this gap through explainable and reliable machine learning approaches tailored to materials discovery. Kailkhura *et al.* [9] advocate for reliable and explainable methods to accelerate materials discovery, arguing that interpretability can enhance user confidence and facilitate debugging. Zhong *et al.* [10] extend this by reviewing explainable ML techniques specifically for materials science, demonstrating how feature importance and counterfactual explanations can make predictions more transparent. Nevertheless, these works focus primarily on technical mechanisms rather than empirical evaluation of whether such explanations actually improve human trust or appropriate reliance.

Broader AI trust literature integrated into materials contexts further underscores the gap. Glikson and Woolley [8] reviewed empirical research on human trust in AI and found that competence, process, and purpose perceptions strongly predict reliance behavior—insights directly applicable yet underutilized in materials AI. Mehrotra *et al.* [12] conducted a systematic review of appropriate trust in human-AI interaction, identifying design strategies that could be adapted to

materials workflows but noting their absence in domain-specific studies. Pataranutaporn *et al.* [13] showed that priming beliefs about AI can enhance perceived trustworthiness, empathy, and effectiveness, suggesting simple interventions that materials AI developers have yet to test systematically.

Quantitative estimates from the reviewed corpus reinforce the observation: of the 30 studies examined, fewer than 20% explicitly mention “trust” or “trustworthiness” in relation to materials AI applications. Even fewer (approximately 10%) evaluate trust through human-subject experiments or propose trust-specific metrics. Papers that do address trustworthiness tend to equate it with robustness or uncertainty quantification [9, 23] without measuring subjective human responses or calibration. DeCost *et al.* [23] discuss scientific AI paradigms for sustainable materials research and call for scalable, trustworthy systems, yet their analysis remains high-level without concrete trust evaluation protocols.

This gap is particularly concerning given the high-stakes nature of materials science. False positives in property prediction can waste synthesis resources, while false negatives may cause promising materials to be overlooked. Distribution shifts—common when moving from computational to experimental data—further erode reliability, yet few studies examine how trust recovers (or fails to recover) after such discrepancies [11, 16]. Moreover, materials AI increasingly supports industry and regulatory decisions, where benevolence and accountability become paramount [20, 21], yet stakeholder-specific trust models are absent.

In summary, while materials AI has made remarkable technical strides, the field has not yet developed a mature conceptual or empirical understanding of trust. The following sections build on this analysis by proposing domain-adapted trust dimensions, articulating open questions, and offering concrete recommendations for moving toward trust-aware materials AI development.

Dimensions of Trust for Materials AI

Building on the conceptual foundations and existing frameworks reviewed earlier, this section proposes six dimensions of trust specifically adapted to the unique demands of materials AI. These dimensions extend general AI trust models by incorporating the high-stakes, data-sparse, and experimentally irreversible nature of materials discovery. A conceptual framework can be visualized as a hexagonal model in which each dimension forms a vertex, with bidirectional arrows illustrating interdependencies. For instance, transparency reinforces predictive competence while robustness moderates the impact of uncertainty calibration on overall trust. Predictive competence serves as the foundational node, feeding into all others, while accountability acts as an overarching integrative element that links back to benevolence and transparency. This interconnected structure emphasizes that trust in materials AI emerges from the dynamic interplay among dimensions rather than isolated attributes.

Figure 2 visualizes trust in materials AI as an interconnected six-dimensional architecture in which appropriate reliance emerges from the combined effects of competence, calibration, transparency, robustness, benevolence, and accountability.

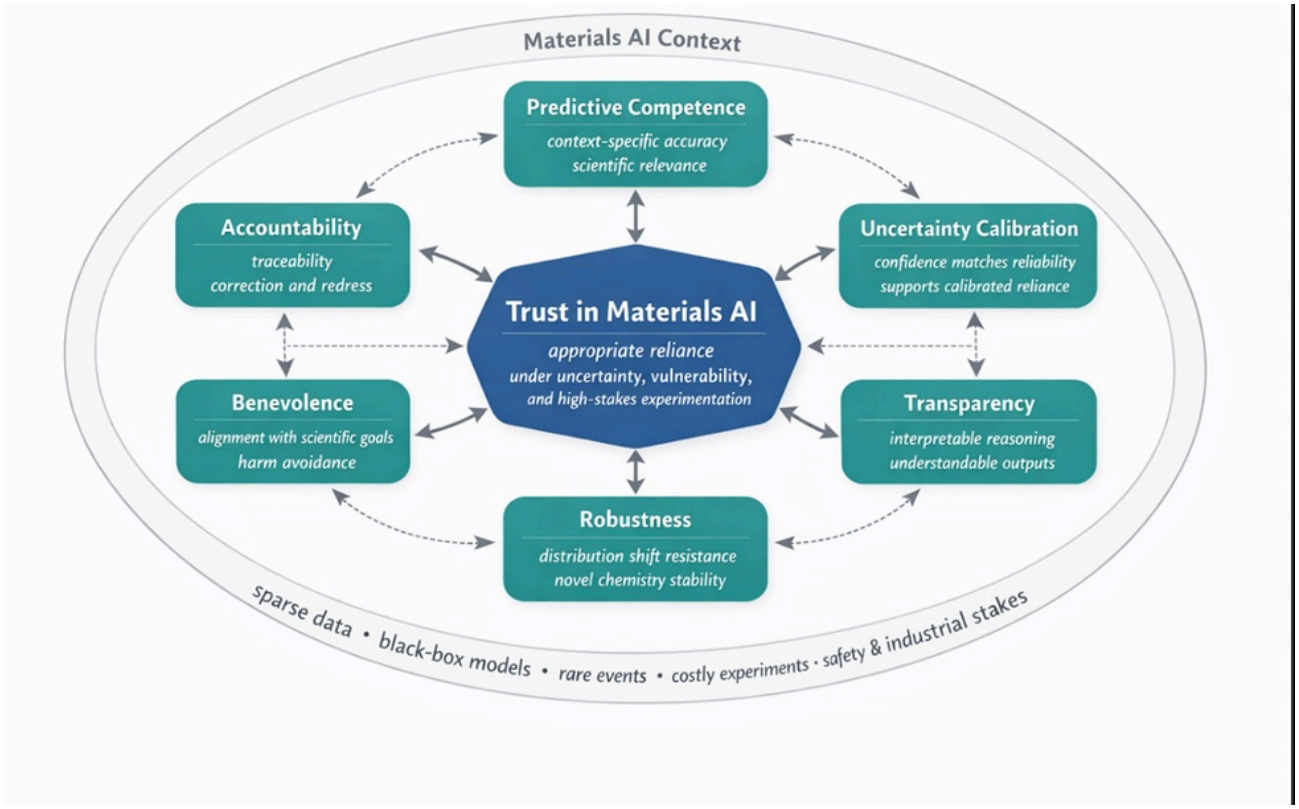

Figure 2. Trust in materials AI as an interconnected six-dimensional architecture in which appropriate reliance emerges from the combined effects of competence, calibration, transparency, robustness, benevolence, and accountability.

Predictive competence refers to the perceived ability of the AI system to generate accurate and relevant predictions within the materials domain. As highlighted by studies on machine learning for molecular and materials science, competence is not merely aggregate accuracy but context-specific performance across chemical spaces [4, 5]. For example, a model predicting formation energies may excel for common oxides yet fail for rare-earth alloys, leading researchers to question its competence even when overall metrics appear strong. Evaluation involves cross-validation on held-out experimental datasets, domain-specific benchmarks, and user studies measuring alignment between predicted and observed outcomes [9, 10]. Without demonstrated competence, other dimensions cannot compensate, as users rapidly disengage from systems perceived as fundamentally unreliable.

Uncertainty calibration concerns the alignment between a model's stated confidence and its actual predictive reliability, a critical factor in materials AI where predictions often guide costly synthesis decisions. Research on reliable machine learning methods underscores that well-calibrated uncertainty estimates enable appropriate reliance, preventing both overconfidence in noisy regions and undue skepticism toward valid but uncertain suggestions [9, 23]. In practice, a materials AI tool recommending a novel perovskite might assign a 95% confidence interval that fails to capture experimental variance, eroding trust. Evaluation requires proper scoring rules such as negative log-likelihood or calibration plots tailored to sparse materials datasets, supplemented by human-subject experiments assessing whether researchers adjust reliance according to reported uncertainties [16, 25].

Transparency denotes the extent to which the internal reasoning of the AI system is accessible and interpretable to domain experts. Multiple studies demonstrate that post-hoc explanations, such as feature attribution or counterfactuals, can enhance perceived trustworthiness in scientific AI contexts [10, 14]. Within materials workflows, transparency might involve visualizing how a graph neural network weighs atomic environments when predicting mechanical properties. However, as noted in broader human-AI interaction reviews, excessive complexity in explanations can paradoxically

reduce trust if users cannot relate them to physical intuition [12, 13]. Evaluation combines objective metrics (e.g., fidelity of explanations to model behavior) with subjective assessments via think-aloud protocols with materials scientists.

Robustness captures the system's capacity to maintain performance and engender trust under distribution shifts, adversarial inputs, or extrapolation to novel chemistries—conditions ubiquitous in materials discovery. Empirical work on explainable and reliable methods for accelerated discovery shows that robustness testing against out-of-distribution samples is essential for sustained user confidence [9, 10, 23]. A materials AI system trained primarily on computational databases may falter when confronted with experimental noise, causing trust to collapse unless robustness is explicitly quantified and communicated. Evaluation includes stress-testing on shifted datasets, adversarial robustness benchmarks, and longitudinal studies tracking trust recovery after observed failures [11, 26-28].

Benevolence reflects the user's perception that the AI system prioritizes human and scientific goals, avoiding recommendations that could lead to harm, wasted resources, or ethical concerns. Frameworks from human-AI collaboration research emphasize that perceived benevolence—manifested through safe default behaviors or explicit harm-avoidance signals—strongly predicts acceptance in high-stakes domains [8, 20, 21]. In materials AI, benevolence might appear when a model flags potentially toxic or unstable candidates rather than simply maximizing predicted performance. Evaluation draws on user surveys measuring alignment with researcher values, as well as scenario-based experiments where participants rate system intent [13, 27].

Accountability addresses the traceability of trust violations and the availability of mechanisms for correction and redress. Studies on justified trust and trustworthiness criteria highlight that accountability closes the loop between system errors and user learning, enabling dynamic trust recalibration [19, 20, 22]. For materials researchers, accountability could involve audit logs linking a failed prediction to specific training data or design choices. Evaluation protocols include auditability checklists, post-mortem analysis frameworks, and metrics quantifying how quickly trust recovers after documented corrections [18, 29, 30].

Table 2 translates the six proposed trust dimensions into an operational evaluation matrix linking each dimension to measurable indicators, characteristic failure signals, and assessment strategies.

Table 2. Evaluation matrix for trust in materials AI: dimensions, indicators, failure signals, and assessment strategies

Trust dimension	Core evaluative question	Technical indicators	Human-centered indicators	Typical trust failure signal	Recommended assessment strategy
Predictive competence	Does the system perform reliably for the relevant materials task and domain region?	Domain-specific benchmark accuracy, external validation, and error stability across chemical spaces	User confidence in recommendation relevance, acceptance of AI-supported candidate ranking	The model appears strong globally, but fails on scientifically important subdomains	Combine held-out experimental validation with task-specific reliance studies
Uncertainty calibration	Do stated confidence levels correspond to actual predictive reliability?	Calibration curves, proper scoring rules, uncertainty-error alignment	User adjustment of reliance according to confidence information	Over-trust in high-confidence wrong outputs or under-trust in low-confidence but useful outputs	Pair formal calibration analysis with human experiments on confidence communication

Transparency	Can domain experts understand why the system produced a recommendation?	Explanation fidelity, feature attribution coherence, counterfactual consistency	Perceived intelligibility, explanation usefulness, and mental-model alignment	Explanations are technically available but scientifically unconvincing or cognitively unusable	Test explanations with think-aloud protocols and expert interpretability tasks
Robustness	Does performance remain credible under shift, novelty, or noisy real-world conditions?	Out-of-distribution testing, stress testing, adversarial or perturbation resilience	Sustained user trust after encountering model limits or environmental change	Trust collapses after failure on novel chemistries or experimental noise	Include longitudinal trust studies after controlled failures and recovery interventions
Benevolence	Does the system appear aligned with scientific, safety, and resource-protection goals?	Safe-default mechanisms, hazard flagging, and conservative recommendation policies	Perceived value alignment, perceived harm avoidance, and acceptance in high-stakes use	Users view the system as optimizing prediction at the expense of safety or scientific judgment	Use scenario-based experiments and value-alignment surveys across stakeholder groups
Accountability	Can errors be traced, explained, corrected, and learned from?	Audit logs, traceability of training data, and model versioning, correction mechanisms	Perceived fairness, confidence in redress, willingness to reuse after correction	Trust fails to recover because users cannot identify why the system was wrong	Require auditability protocols and post-failure correction pathways in evaluation pipelines

Collectively, these six dimensions form a comprehensive, materials-specific trust model that addresses the limitations of generic frameworks. They shift focus from isolated technical improvements to holistic socio-technical design, ensuring that materials AI systems not only perform but also earn and maintain human trust.

Open Questions and Future Directions

Despite the conceptual advances outlined above, several critical open questions remain regarding trust in materials AI, each pointing toward promising research directions. These questions emerge directly from the reviewed corpus and highlight the interdisciplinary gaps between human factors, AI trustworthiness engineering, and materials-specific applications.

How is trust initially established in the first-time use of a materials AI system? Studies on dispositional and situational trust indicate that initial impressions are shaped by interface cues and prior beliefs about automation [8, 13, 24]. Yet, materials researchers often encounter novel tools with little domain-specific priming. Future work could employ controlled onboarding experiments to quantify how tutorial design, uncertainty visualizations, and performance demonstrations influence baseline trust levels.

How does trust evolve after failures such as false positives or false negatives in materials predictions? Dynamic trust models suggest asymmetric recovery patterns, with negative outcomes eroding trust more rapidly than positive ones

restore it [2, 11, 16]. Longitudinal field studies tracking materials scientists' reliance behavior across repeated interactions with autonomous discovery platforms are urgently needed to map these trajectories empirically.

What is the precise relationship between explainability techniques and actual trust calibration in materials contexts? While transparency is frequently proposed as a trust enhancer [10, 14], some evidence indicates that certain explanations may increase over-reliance on flawed models [12, 26]. Targeted experiments comparing different explanation modalities (e.g., SHAP versus counterfactuals) within materials property prediction tasks would clarify optimal design choices.

How should trust be calibrated for rare-event predictions, such as the discovery of entirely novel materials with no close analogs in training data? The reviewed literature on uncertainty calibration and robustness reveals limited guidance for extrapolation scenarios [9, 23, 25], where traditional metrics break down. Future research might develop specialized calibration protocols that incorporate physics-informed priors or active learning feedback loops.

How does trust in materials AI differ across stakeholder groups, including academic researchers, industrial practitioners, and regulatory bodies? Preliminary findings from broader AI trust studies suggest stakeholder-specific weighting of dimensions such as benevolence and accountability [19, 21, 27], yet materials-specific comparative analyses are absent. Multi-stakeholder surveys and ethnographic studies could illuminate these variations and inform tailored interface designs.

Can trust be measured objectively through behavioral or physiological markers, or does it remain inherently subjective? Emerging work on psychophysiological responses and machine learning-based trust prediction offers preliminary pathways [15, 28], but validation within materials workflows remains exploratory. Hybrid measurement frameworks combining self-reports, eye-tracking, and interaction logs represent a key methodological frontier.

Additional open questions concern the long-term societal implications of pervasive materials AI trust, including potential deskilling of human expertise and the ethical governance of autonomous discovery systems [20, 22]. Addressing these questions will require sustained collaboration across human-computer interaction, psychology, and materials science communities, moving beyond accuracy-only benchmarks toward comprehensive trust-aware evaluation protocols.

Recommendations

To translate the conceptual insights of this review into practice, the following stakeholder-specific recommendations are offered.

For authors developing materials AI systems: (a) explicitly state assumptions regarding user trust and intended reliance levels within technical publications, moving beyond accuracy reporting to include dimension-specific trustworthiness metrics [9, 10, 23]; (b) incorporate human-subject evaluations of trust calibration alongside model performance benchmarks, using protocols adapted from human-AI interaction studies [12, 13, 24]; (c) report uncertainty calibration and robustness results transparently, including failure modes relevant to real-world materials workflows [16, 25]; and (d) design interfaces that communicate benevolence and accountability features proactively, such as safety flags or correction mechanisms [20, 21].

For reviewers evaluating manuscripts on materials AI: (a) require explicit discussion of how proposed methods address at least three trust dimensions, rejecting accuracy-only claims [4, 5, 8]; (b) question implicit over-trust assumptions by demanding evidence of appropriate reliance testing rather than assuming user acceptance [2, 18]; and (c) insist on calibration evidence and stakeholder considerations, elevating trust evaluation to the level of methodological rigor expected for predictive performance [19, 22].

For the broader materials AI community: (a) establish standardized trust benchmarks for common tasks such as property prediction and inverse design, analogous to existing materials benchmarking initiatives [7, 23]; (b) develop and

disseminate open-source trust evaluation protocols that integrate the six proposed dimensions into existing ML pipelines [9, 10, 30]; and (c) initiate interdisciplinary research programs examining trust dynamics within actual laboratory and industrial materials workflows, fostering empirical datasets on human-AI collaboration [11, 27, 28].

These recommendations, if adopted, would accelerate the maturation of materials AI from technically impressive tools to genuinely trustworthy partners in discovery. Implementation will benefit from community-driven initiatives, including workshops and shared repositories, to ensure collective progress toward trust-aware development.

Conclusion

This review has synthesized conceptual models of trust from human factors, psychology, and human-computer interaction with the current landscape of materials AI. It demonstrates that while predictive capabilities have advanced rapidly, the field has largely overlooked the socio-technical dimensions essential for effective human-AI collaboration in high-stakes discovery contexts. By proposing six interconnected dimensions—predictive competence, uncertainty calibration, transparency, robustness, benevolence, and accountability—the work provides a practical framework for designing, evaluating, and deploying trustworthy materials AI systems. The identified open questions underscore critical gaps in initial trust formation, post-failure dynamics, and stakeholder variations. At the same time, the recommendations offer actionable pathways for authors, reviewers, and the community to bridge these gaps.

Ultimately, moving beyond accuracy-centric evaluation toward trust-aware design, evaluation, and reporting is not merely an incremental improvement but a necessary paradigm shift. Only by embedding rigorous conceptual models of trust can materials AI fulfill its promise of accelerating innovation while safeguarding scientific integrity and human judgment. Future research guided by the frameworks and questions articulated here will ensure that materials discovery remains a collaborative, reliable, and ethically grounded endeavor between humans and intelligent systems.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 17 Dec 2023 Revised: 17 Feb 2024 Accepted: 26 Apr 2024
Published online: 18 July 2024

References

Guo W. Explainable artificial intelligence for 6G: Improving trust between human and machine. IEEE Commun Mag. 2020;58(6):39-45.

Rodriguez L, Bustamante Orellana CE, Chiou EK, Huang L, Cooke N, Kang Y. A review of mathematical models of human trust in automation. Front Neuroergon. 2023;4:1171403.

Love PE, Matthews J, Fang W, Porter S, Luo H, Ding L. Learning to comprehend and trust artificial intelligence outcomes: A conceptual explainable AI evaluation framework. IEEE Eng Manag Rev. 2023;52(1):230-47.

Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nature. 2018;559(7715):547-55.

Schmidt J, Marques MR, Botti S, Marques MA. Recent advances and applications of machine learning in solid-state materials science. npj Comput Mater. 2019;5(1):83.

Zunger A. Inverse design in search of materials with target functionalities. Nat Rev Chem. 2018;2(4):0121.

Montoya JH, Aykol M, Anapolsky A, Gopal CB, Herring PK, Hummelshøj JS, et al. Toward autonomous materials research: Recent progress and future challenges. Appl Phys Rev. 2022;9(1):011405.

Glikson E, Woolley AW. Human trust in artificial intelligence: Review of empirical research. Acad Manag Ann. 2020;14(2):627-60.

Kailkhura B, Gallagher B, Kim S, Hiszpanski A, Han TY. Reliable and explainable machine-learning methods for accelerated material discovery. npj Comput Mater. 2019;5(1):108.

Zhong X, Gallagher B, Liu S, Kailkhura B, Hiszpanski A, Han TY. Explainable machine learning in materials science. npj Comput Mater. 2022;8(1):204.

Edmonds M, Gao F, Liu H, Xie X, Qi S, Rothrock B, et al. A tale of two explanations: Enhancing human trust by explaining robot behavior. Sci Robot. 2019;4(37):eaay4663.

Mehrotra S, Degachi C, Vereschak O, Jonker CM, Tielman ML. A systematic review on fostering appropriate trust in human-AI interaction: Trends, opportunities and challenges. ACM J Responsib Comput. 2024;1(4):1-45.

Pataranutaporn P, Liu R, Finn E, Maes P. Influencing human–AI interaction by priming beliefs about AI can increase perceived trustworthiness, empathy and effectiveness. Nat Mach Intell. 2023;5(10):1076-86.

Göbel K, Niessen C, Seufert S, Schmid U. Explanatory machine learning for justified trust in human-AI collaboration: Experiments on file deletion recommendations. Front Artif Intell. 2022;5:919534.

Li Z, Lu Z, Yin M. Modeling human trust and reliance in AI-assisted decision making: A markovian approach. In: Proceedings of the AAAI Conference on Artificial Intelligence. 2023;37(5):6056-64.

Roeder L, Hoyte P, van der Meer J, Fell L, Johnston P, Kerr G, et al. A quantum model of trust calibration in human–AI interactions. Entropy. 2023;25(9):1362.

Choung H, David P, Ross A. Trust in AI and its role in the acceptance of AI technologies. Int J Hum Comput Interact. 2023;39(9):1727-39.

Wischnewski M, Krämer N, Müller E. Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. 2023:1-16.

Kostick-Quenet K, Lang BH, Smith J, Hurley M, Blumenthal-Barby J. Trust criteria for artificial intelligence in health: Normative and epistemic considerations. J Med Ethics. 2024;50(8):544-51.

Ferrario A. Justifying our credences in the trustworthiness of AI systems: A reliabilistic approach. Sci Eng Ethics. 2024;30(6):55.

Afroogh S, Akbari A, Malone E, Kargar M, Alambeigi H. Trust in AI: Progress, challenges, and future directions. Humanit Soc Sci Commun. 2024;11(1):1568.

Lukyanenko R, Maass W, Storey VC. Trust in artificial intelligence: From a foundational trust framework to emerging research opportunities. Electron Mark. 2022;32(4):1993-2020.

DeCost BL, Hattrick-Simpers JR, Trautt Z, Kusne AG, Campo E, Green ML. Scientific AI in materials science: A path to a sustainable and scalable paradigm. Mach Learn Sci Technol. 2020;1(3):033001.

Ueno T, Sawa Y, Kim Y, Urakami J, Oura H, Seaborn K. Trust in human-AI interaction: Scoping out models, measures, and methods. In: CHI Conference on Human Factors in Computing Systems Extended Abstracts. 2022:1-7.

Love J, Gronau QF, Palmer G, Eidels A, Brown SD. In human–machine trust, humans rely on a simple averaging strategy. Cogn Res Princ Implic. 2024;9(1):58.

Lee E. How do we build trust in machine learning models? SSRN Electron J. 2021. https://doi.org/10.2139/ssrn.3822437.

Li M, Kamaraj AV, Lee JD. Modeling trust dimensions and dynamics in human-agent conversation: A trajectory epistemic network analysis approach. Int J Hum Comput Interact. 2024;40(14):3571-82.

Chauhan H, Jang Y, Jeong I. Predicting human trust in human-robot collaborations using machine learning and psychophysiological responses. Adv Eng Inform. 2024;62:102720.

Naiseh M, Al-Thani D, Jiang N, Ali R. Explainable recommendation: When design meets trust calibration. World Wide Web. 2021;24(5):1857-84.

Rattan AK. Data integrity: History, issues, and remediation of issues. PDA J Pharm Sci Technol. 2018;72(2):105-16.