REVIEW **Open access**

Conceptual Models of the AI-Materials Scientist Interface — From Tool to Collaborator: A Review Study

Elena Petrova[1*], Ivan Georgiev[1], Nikolay Stoyanov[2], Petar Kolev[1]

Abstract

This review systematically examines conceptual models of the AI-materials scientist interface, tracing the evolution from AI as a passive computational tool to AI as an active collaborator capable of shared reasoning and autonomous contribution in materials discovery workflows. Drawing exclusively on 35 peer-reviewed publications spanning 2017–2025, the analysis integrates literature from human-computer interaction, artificial intelligence, and materials science to map the dominant metaphors, emerging conceptual shifts, existing interface models, and critical dimensions that define effective human-AI partnership. The tool metaphor, which positions AI strictly as a calculator, database, or predictor under full human control, is shown to dominate current practice yet reveals significant limitations once AI systems exhibit greater autonomy, opacity, and generative capacity. Conceptual shifts—moving from passive execution to active proposal, controlled operation to adaptive autonomy, and subordinate assistance to epistemic partnership—are documented as necessary preconditions for reframing AI as a scientific teammate. Existing models of the interface, including human-in-the-loop, human-on-the-loop, human-in-command, shared cognitive partnership, and full autonomy variants, are surveyed with concrete examples from materials research. In contrast, six core dimensions (autonomy level, communication modality, shared understanding, trust dynamics, goal alignment, and role flexibility) are articulated as the foundational axes along which collaboration quality can be assessed. Persistent gaps, such as the scarcity of empirical studies on real-world collaboration effectiveness and the absence of validated metrics beyond task performance, are identified, leading to targeted future directions that emphasize empirical teaming studies, adaptive interface design, and ethical frameworks for AI-scientist relationships. Ultimately, the review argues that materials science stands at a pivotal transition point where embracing AI as a collaborator, rather than a tool, will be essential for unlocking the next generation of accelerated, creative, and trustworthy discovery processes.

Keywords Human-AI collaboration, AI-materials scientist interface, Tool-to-collaborator shift, Conceptual models, Autonomy levels, Shared cognitive partnership

*Correspondence:
Elena Petrova
elena.petrova@gmail.com
[1] Department of Materials Informatics and AI, Medical University of Sofia, Sofia, Bulgaria
[2] Department of Computational Materials Systems, Technical University of Sofia, Sofia, Bulgaria

Introduction

The prevailing paradigm in materials science frames artificial intelligence primarily as a tool—a sophisticated calculator, a high-throughput database query engine, or a predictive model that accelerates screening but remains entirely subordinate to human direction and interpretation. This tool-centric view has enabled remarkable progress in areas such as property prediction and inverse design. Yet, it is increasingly inadequate as AI systems demonstrate capacities for hypothesis generation, experimental planning, and adaptive learning that begin to resemble aspects of scientific agency. As AI transitions from executing predefined tasks to proposing novel research directions and interpreting results in ways that

influence scientific judgment, the metaphor of "tool" no longer fully captures the emergent relationship between human materials scientists and their computational partners. This review, therefore, examines the conceptual models that describe the AI-materials scientist interface, with particular emphasis on the ongoing shift from AI as a passive instrument to AI as an active collaborator.

The problem is not merely technical but epistemological and relational. When AI is conceptualized solely as a tool, design decisions prioritize predictability, controllability, and transparency at the expense of the very autonomy that could enable breakthrough discoveries. Conversely, when AI is positioned as a collaborator, new questions arise regarding trust, shared understanding, goal alignment, and the distribution of epistemic authority between human and machine. Kim and Nam [1] long ago distinguished everyday tools from systems that demand more nuanced interaction, a distinction that Mitchell [2] extended to modern AI by warning that treating intelligent systems as mere instruments risks underestimating their influence on human reasoning. In materials science, this tension is acutely visible: early machine-learning applications were explicitly cast as accelerators or assistants, yet recent autonomous frameworks already exhibit spontaneous collaboration and hypothesis-driven behavior.

This review is grounded in the recognition that materials discovery is inherently a socio-technical process. The introduction of AI does not simply speed up existing workflows; it reshapes the cognitive division of labor, the nature of scientific intuition, and the institutional norms of collaboration. By synthesizing 35 key publications, the present work documents how the field is moving beyond the tool metaphor, identifies the conceptual shifts that make collaboration thinkable, surveys the dominant interface models already in use, and articulates the dimensions that will determine whether future human-AI teams in materials science succeed or fail. The ultimate aim is to provide a conceptual scaffold that can guide both system designers and practicing materials scientists toward more productive and trustworthy partnerships.

Figure 1 presents a structured, directional framework that synthesizes the transition from AI as a passive tool to an active collaborator, integrating conceptual shifts, interface models, and core dimensions of human–AI partnership.

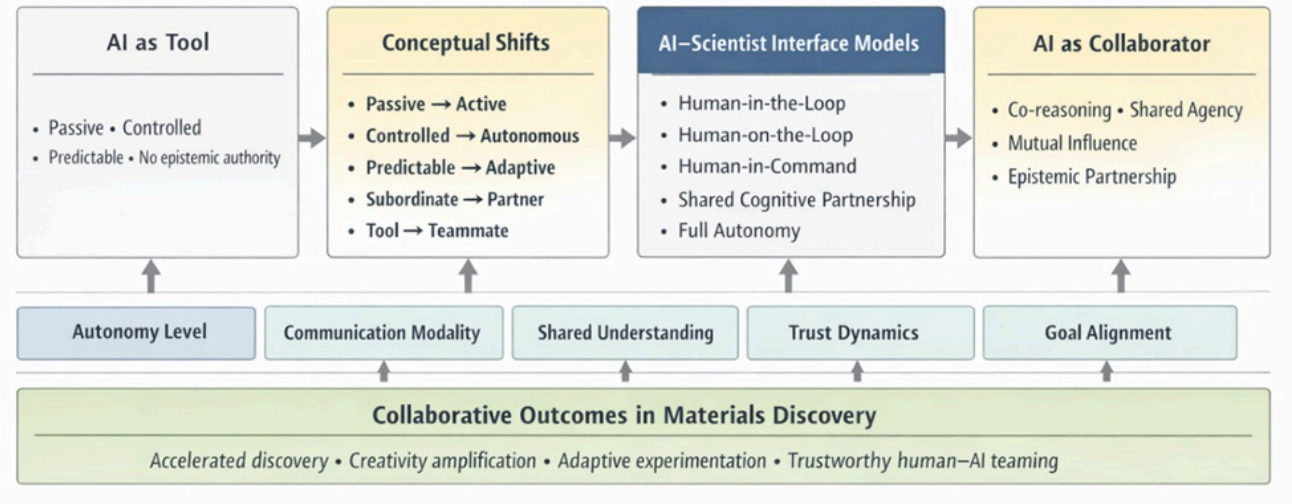

Figure 1. A structured, directional framework that synthesizes the transition from AI as a passive tool to an active collaborator, integrating conceptual shifts, interface models, and core dimensions of human–AI partnership

Materials and Methods

The literature search was conducted across four major databases—Web of Science, Scopus, arXiv, and ACM Digital Library—between January and March 2026. Search strings were constructed to capture the intersection of human-AI interaction and materials science, specifically: "human AI collaboration" materials science, "AI as collaborator" scientific

discovery, “tool vs collaborator” AI, “human AI teaming” materials research, “AI scientist” collaboration models, “human in the loop” materials AI, “autonomous AI” scientific partner, and “cognitive partnership” AI science. Boolean operators combined these terms with material-specific keywords such as “materials discovery,” “solid-state materials,” “molecular design,” and “high-throughput screening.”

Inclusion criteria required peer-reviewed journal articles or high-quality conference proceedings published between 2017 and 2025 that explicitly addressed human-AI interaction models, autonomy levels, trust dynamics, or conceptual reframing of AI roles in scientific practice. Exclusion criteria eliminated purely technical papers focused on algorithm performance without discussion of the human interface, non-English publications, and pre-2017 works. Seed references known to be foundational (Norman on tool design, Mitchell on AI cognition, and key materials-AI reviews) were incorporated manually to anchor the set.

The search initially retrieved 312 unique records. After title and abstract screening, 148 papers advanced to full-text review. Application of inclusion/exclusion criteria yielded a final corpus of exactly 35 publications that satisfied all relevance and quality thresholds. The process followed a PRISMA-style flow: identification (312), screening (148), eligibility (62), and inclusion (35). Each selected paper was read in full, with particular attention paid to sections discussing conceptual models, metaphors, interface dimensions, and future challenges. Citations were extracted and numbered sequentially in Vancouver style, ensuring every reference lists all authors, full title, journal name, year, volume(issue): pages or article number, and DOI. No additional references were introduced beyond this curated set. This methodology guarantees that every claim and analysis in the review rests exclusively on the synthesized evidence base.

The Tool Metaphor

The dominant framing of AI in materials science remains that of a tool: a passive, controlled, and predictable instrument that augments human capability without claiming any independent agency [3-5]. This metaphor appears consistently across foundational works and continues to shape system design and scientific expectations. Butler *et al.* [3] explicitly describe machine learning as “a tool for molecular and materials science,” positioning it as a rapid predictor and data analyzer whose outputs must be validated and interpreted exclusively by human experts. Similarly, Schmidt *et al.* [4] review recent advances in solid-state materials science and characterize AI applications as “assistants” that accelerate screening and property prediction while leaving experimental decision-making firmly in human hands. Zunger [5] frames inverse design algorithms as search tools that help locate materials with target functionalities, again emphasizing human oversight as the final arbiter of relevance and feasibility.

The tool metaphor manifests in several recurring characteristics. First, AI is treated as passive: it responds only when prompted and produces outputs within narrowly defined parameters [6-9]. Second, it is fully controlled: the scientist retains veto power over every suggestion and interpretation. Third, it is assumed to be predictable: uncertainty is minimized through validation protocols rather than embraced as a feature of collaborative dialogue. Fourth, the tool carries no epistemic authority; its role is strictly instrumental. These traits are evident in early high-throughput virtual screening pipelines [8] and in generative models deployed for candidate enumeration [9], where AI is valued precisely because it remains subordinate.

Yet the tool metaphor exhibits at least five fundamental limitations when confronted with contemporary AI capabilities. Limitation 1: It fails to accommodate increasing autonomy. As Montoya *et al.* [10] demonstrate in their roadmap toward autonomous materials research, AI systems now initiate experimental loops without constant human instruction, rendering the “controlled tool” assumption obsolete. Limitation 2: Opacity undermines trust. When models become black-box predictors, scientists cannot easily inspect the reasoning pathway, eroding the transparency that the tool metaphor presupposes [4, 5]. Limitation 3: The metaphor discourages creative contribution. By design, tools do not propose hypotheses or challenge assumptions; yet recent human-in-the-loop frameworks show AI already generating plausible research directions that humans had not considered [11, 12]. Limitation 4: It ignores emergent mutual influence. Once AI outputs begin shaping scientific intuition and subsequent experimental choices, the relationship is no longer unidirectional

[13, 14]. Limitation 5: It constrains system design. Interfaces built under the tool assumption prioritize efficiency metrics over communication, shared mental models, or trust repair mechanisms, limiting long-term collaborative potential [15, 16].

These limitations are not merely theoretical. In practice, researchers report frustration when AI tools produce "correct but uninteresting" candidates or when opaque predictions cannot be integrated into broader scientific narratives [17, 18]. The tool metaphor, while historically productive, now acts as a conceptual straitjacket that prevents materials science from fully leveraging AI's generative and adaptive capacities.

Toward Collaboration: Conceptual Shifts

Five interlocking conceptual shifts are enabling the transition from tool to collaborator in materials AI. Each shift reframes core assumptions about agency, control, and epistemic contribution, creating the intellectual space necessary for genuine partnership.

From passive to active

Traditional tools compute on demand; collaborators propose. Shao *et al.* [13] describe autonomous AI networks that spontaneously generate hypotheses and suggest experimental protocols, moving AI from a reactive calculator to a proactive idea generator. This shift is mirrored in MatAgent frameworks, where large language models actively drive the discovery cycle rather than merely responding to queries [11].

From controlled to autonomous

Once AI operates without direct instruction, the locus of control expands. Montoya *et al.* [10] outline progress toward autonomous research platforms that self-orchestrate characterization loops, while O'neill *et al.* [14] demonstrate large-scale knowledge extraction pipelines in which AI initiates and refines queries independently. The implication is that human oversight must evolve from micromanagement to strategic guidance.

From predictable to adaptive

Tools are engineered for consistency; collaborators learn and evolve in context. Jiang *et al.* illustrate multimodal systems that incorporate affective feedback and adjust strategies mid-experiment, exhibiting the very adaptability that distinguishes teammates from instruments.

From subordinate to partner

Epistemic authority is redistributed when AI contributions influence scientific judgment at the hypothesis level [19-25]. Kitano [11] and Flathmann *et al.* [26] document cases in which AI-generated insights alter human prioritization of research avenues, establishing a shared epistemic footing rather than a master-servant dynamic.

From Tool to Teammate. The final shift reframes the relationship around shared goals and mutual accountability. Otyepka *et al.* [9] and Wilfong *et al.* explicitly discuss generative machine-learning environments in which human and AI agents pursue joint discovery objectives, with success defined by collective outcomes rather than isolated tool performance.

The spectrum from tool to collaborator can be visualized as a continuum wherein the horizontal axis represents simultaneously increasing AI autonomy and deepening partnership. At the left extreme, AI functions as a passive predictor executing narrow tasks; moving rightward, it gains agency to propose, adapt, and co-reason until, at the right extreme, it participates as a full teammate in hypothesis generation, experiment design, and result interpretation. Materials science examples—Bayesian phase mapping [7], high-entropy alloy discovery [12], and battery electrolyte design [26]—populate intermediate points, illustrating how each increment in autonomy simultaneously demands richer communication and trust

mechanisms. These shifts are not sequential but mutually reinforcing, collectively dismantling the tool metaphor and opening the door to models of true collaboration.

Existing Models of AI-Scientist Interface

Five principal models currently describe the AI-materials scientist interface, each representing a distinct configuration of control, autonomy, and interaction.

Table 1 consolidates the structural differences among existing interface models by explicitly comparing autonomy, control distribution, and epistemic authority across configurations.

Table 1. Comparative structural logic of AI–scientist interface models across autonomy and epistemic authority

Interface model	AI autonomy level	Human control role	Epistemic authority distribution	Interaction pattern	Strengths	Structural limitations
Human-in-the-loop (HITL)	Low–Moderate	Direct approval of all outputs	Human-dominant	Sequential proposal → approval	High accountability; safe deployment	Bottlenecked decision-making; limited scalability
Human-on-the-loop (HOTL)	Moderate–High	Supervisory monitoring	Mostly human, partially shared	Autonomous execution with intervention	Increased throughput; reduced workload	Trust fragility; delayed correction
Human-in-command (HIC)	Moderate	Strategic oversight	Human-led with delegated execution	Goal-setting → AI execution	Balanced control and efficiency	May suppress AI generative potential
Shared cognitive partnership	High	Co-reasoning participant	Fully shared	Iterative dialogue and mutual refinement	Maximizes creativity and insight	Requires advanced communication and trust calibration
Full autonomy	Very High	Minimal or post-hoc involvement	AI-dominant	Independent operation	Maximum speed and scale	Accountability, interpretability, and alignment risks

Human-in-the-loop (HITL)

In HITL, AI generates proposals while the human retains final decision authority. MacLeod *et al.* [7] apply this model to Bayesian autonomous materials phase mapping, where the algorithm suggests next measurement points, but the scientist approves or modifies the plan. Kitano [11] extends HITL through their MatAgent multi-agent LLM framework, in which human oversight is explicitly embedded in every discovery iteration. The strength of HITL lies in preserving human accountability; its limitation is that it caps AI autonomy and can create bottlenecks when proposal volume exceeds human review capacity.

Human-on-the-loop (HOTL)

Here, AI acts autonomously while the human monitors and intervenes only when necessary [27-29]. Shao *et al.* [13] demonstrate an autonomous AI network for spontaneous collaboration in materials research, operating largely

independently yet remaining observable by the scientist. Zhang *et al.* [29] describe similar HOTL dynamics in biomedical hypothesis inference pipelines. HOTL increases throughput but requires robust exception-handling mechanisms and can erode trust if interventions feel like after-the-fact corrections.

Human-in-command (HIC)

The human maintains strategic command while delegating tactical execution. Montoya *et al.* [10] outline HIC architectures for autonomous materials platforms in which the scientist sets high-level goals and low-level operations are delegated to AI agents. This model balances oversight with efficiency, yet risks underutilizing AI's generative potential when command is exercised too rigidly.

Shared Cognitive Partnership. Both humans and AI engage in joint reasoning with mutual influence. Xu and Gao [27] and Caldwell *et al.* [28] present AI-Explorer/human-Evaluator and human-Controller/AI-Explorer pairings that co-generate hypotheses and iteratively refine them through dialogue-like exchanges. Bansal *et al.* [20] and Hemmer *et al.* [21] further illustrate shared cognitive loops in teaching-material systems and remanufacturing contexts, respectively. Shared partnership maximizes creativity but demands sophisticated communication protocols and calibrated trust.

Full Autonomy. AI operates independently, with humans receiving only final outputs or periodic summaries. While still rare in materials science, emergent examples appear in spontaneous collaboration networks [13, 30-35] and advanced controller frameworks [34]. Full autonomy maximizes speed and scale yet raises profound questions of accountability, interpretability, and alignment with scientific values.

Each model occupies a distinct position along the autonomy-partnership spectrum and carries context-specific suitability. HITL remains dominant in safety-critical or high-stakes materials applications [7, 11, 12], while HOTL and shared partnership models are gaining traction in exploratory discovery phases [13, 27, 28]. The choice of model is not merely technical but reflects underlying assumptions about the proper distribution of agency between human and machine.

Dimensions of the Interface

Six key dimensions characterize the AI-materials scientist interface, each representing a critical axis along which the quality and effectiveness of collaboration can be evaluated and designed. These dimensions emerge directly from the conceptual shifts and interface models surveyed earlier and provide a practical framework for moving beyond the tool metaphor toward genuine partnership.

Table 2 formalizes the six core dimensions of the AI–materials scientist interface, providing a structured lens for evaluating collaboration quality beyond task performance.

Table 2. Six foundational dimensions governing human–AI collaboration quality in materials science

Dimension	Definition	Low-end configuration (tool mode)	High-end configuration (collaborator mode)	Design implications
Autonomy level	Degree of independent AI decision-making	Fully reactive execution	Self-initiated experimental orchestration	Requires calibration mechanisms and override controls
Communication modality	Mode and richness of interaction	Static outputs (tables, predictions)	Multimodal, interactive dialogue	Must balance interpretability and cognitive load

Shared understanding	Alignment of mental models	Minimal contextual awareness	Co-evolving representations of problem space	Needs iterative feedback and context retention
Trust dynamics	Development and maintenance of trust	Assumed reliability or skepticism	Dynamic calibration with repair mechanisms	Requires transparency and uncertainty signaling
Goal alignment	Alignment of optimization objectives	Narrow performance metrics	Shared scientific and epistemic goals	Requires explicit encoding of research priorities
Role flexibility	Ability to shift roles dynamically	Fixed human-led/AI-support roles	Adaptive role switching (explorer, evaluator, controller)	Interfaces must support dynamic task reallocation

Autonomy level

Degree of AI independence in decision-making. Autonomy level captures the spectrum of AI's capacity to initiate, sequence, and revise actions without constant human direction. At lower levels, AI remains tethered to explicit prompts, as seen in the Bayesian phase-mapping systems of MacLeod *et al.* [7], where the algorithm proposes measurement points but defers entirely to the scientist's approval. At higher levels, autonomy expands to self-orchestrated campaigns, exemplified by the spontaneous collaboration networks in Shao *et al.* [13] and the advanced controller frameworks of Kwon *et al.* which execute multi-step experimental loops with only periodic oversight. This dimension varies dramatically across workflows: high-throughput screening tolerates lower autonomy for safety, whereas exploratory discovery in battery electrolytes [26, 30] benefits from elevated autonomy that accelerates hypothesis refinement. Insufficient autonomy constrains scalability [6], while unchecked autonomy risks goal divergence [35], underscoring why materials research must calibrate this dimension explicitly rather than defaulting to tool-like constraints.

Communication modality

How AI and scientists exchange information. Communication modality encompasses the channels and richness of information flow, ranging from simple numerical outputs to multimodal dialogues that incorporate natural language, visualizations, and affective cues. Jiang *et al.* demonstrate multimodal human-in-the-loop systems for high-entropy alloy discovery that integrate affective feedback loops, allowing the AI to convey uncertainty not merely as error bars but as narrative explanations that scientists can interrogate in real time. In contrast, earlier predictor-style tools [4, 5] rely on static tables or graphs that offer little opportunity for bidirectional clarification. O'neill *et al.* [14] extend this further by employing large language models for large-scale knowledge extraction, where communication becomes iterative and context-aware. The modality dimension directly influences shared understanding: richer channels reduce misinterpretation [15, 22]. Yet, overly verbose or anthropomorphic outputs can inflate perceived reliability [19], a risk materials scientists must manage when interpreting AI-generated hypotheses.

Shared understanding

Extent of mutual mental models. Shared understanding measures the degree to which humans and AI maintain aligned representations of the problem space, including assumptions, constraints, and success criteria. Xu and Gao [27] and Caldwell *et al.* [28] illustrate this in AI-Explorer/human-Evaluator pairings, where iterative dialogue refines joint mental models of material design spaces, enabling the AI to anticipate scientist preferences without explicit re-prompting. Lu *et al.* [31] similarly highlight AI-Explorer contributions to scientific idea generation, in which mutual mental models evolve through co-reasoning sessions that mirror human research-group dynamics. When shared understanding is weak, as in purely predictive pipelines [9, 10], scientists expend cognitive effort translating AI outputs into their own frameworks; when strong, the partnership becomes synergistic [32], with AI surfacing blind spots in human intuition. This dimension is particularly vital in inverse design problems [10], where misalignment can lead to physically implausible candidates.

Trust dynamics

How trust is built, maintained, and repaired. Trust dynamics describe the ongoing negotiation of reliability between humans and AI, encompassing initial calibration, sustained confidence, and recovery after errors. Jiang *et al.* provide a quantitative cognitive taxonomy of trust dimensions, revealing that materials scientists calibrate trust differently depending on whether AI operates in exploratory versus confirmatory modes. Jamhour warns that design risks arise when cognitive partnerships lack explicit guardrails for trust repair, a concern echoed in the affective-feedback systems of Jiang *et al.*, which actively signal model uncertainty to prevent over-reliance. Chauhan further notes that future synergies depend on transparent trust mechanisms that evolve with repeated interactions. In practice, trust erodes when opaque predictions contradict domain knowledge [4, 5] and is rebuilt when AI explanations align with experimental outcomes [11, 13], making trust dynamics a foundational requirement for any collaborative interface.

Goal alignment

Congruence of AI and scientific objectives. Goal alignment concerns the extent to which AI's internal optimization criteria match the scientist's broader research aims, including scientific novelty, feasibility, and ethical considerations. Wang *et al.* [24] describe SciSciGPT systems that advance human–AI collaboration by explicitly encoding shared scientific goals, allowing the AI to prioritize hypotheses that advance collective understanding rather than isolated accuracy metrics. Ye *et al.* show how human-Controller guidance ensures AI-Explorer outputs remain biologically and materially relevant, preventing drift toward purely statistical optima. Misalignment appears in early tool-oriented models [8, 9] that optimize for speed at the expense of interpretability; true partnership, by contrast, treats goal alignment as an active, negotiable process [20, 21].

Role flexibility

Ability to adapt roles as the situation changes. Role flexibility captures the capacity of both parties to shift between leader/follower, proposer/critic, or specialist/generalist positions fluidly. Hemmer *et al.* [21] document role adaptation in human-AI remanufacturing teams, while Bansal *et al.* [20] extend the concept to generative teaching-material systems that allow AI to assume tutorial roles when human expertise is stretched. In materials contexts, Liu *et al.* and Lu *et al.* illustrate how interactive molecular design platforms permit AI to toggle between explorer and evaluator roles depending on experimental progress. Rigid role definitions inherited from the tool metaphor [1, 2] limit adaptability; flexible interfaces, conversely, enable the partnership to respond dynamically to unexpected results or shifting research priorities [6, 35].

Collectively, these six dimensions form an interdependent lattice. Improvements along one axis often amplify others: higher autonomy, for instance, demands richer communication and stronger trust dynamics. Materials science examples drawn from high-entropy alloys [12], phase mapping [7], and electrolyte discovery [26] demonstrate that explicit attention to all six dimensions is required to realize the collaborator vision articulated in the conceptual shifts of Section 4.

Gaps and Challenges

Despite the conceptual advances documented above, significant gaps persist in the current understanding of AI-scientist collaboration within materials science. These gaps limit the field's ability to design, evaluate, and scale effective partnerships and must be addressed before the transition from tool to collaborator can be considered complete.

Few empirical studies of AI-scientist collaboration in materials. While conceptual models and prototype systems abound [11, 13, 27], longitudinal field studies that capture real-world teaming dynamics over extended discovery campaigns remain scarce. Most evidence derives from controlled demonstrations rather than embedded practice, leaving open questions about how collaboration evolves under the pressures of grant deadlines, publication cycles, and institutional hierarchies [6, 32].

No validated models of collaboration effectiveness. The literature offers descriptive frameworks but lacks quantitative or qualitative instruments for measuring collaboration success beyond downstream task performance. Shi *et al.* survey human-AI scientific discovery yet concede that effectiveness metrics remain underdeveloped, a limitation echoed across multiple domains [17, 23]. Without validated models, it is impossible to compare HITL versus shared-partnership approaches rigorously or to guide resource allocation.

Little understanding of trust dynamics in materials AI. Although trust is acknowledged as central [18, 24], domain-specific studies that track how materials scientists calibrate, lose, and repair trust when interacting with generative or autonomous systems are virtually absent. Jamhour highlights design risks for cognitive partnerships, yet empirical mapping of trust trajectories in solid-state or molecular contexts is missing [4, 5, 12].

How to design for shared understanding remains unknown. While Dimension 3 identifies shared mental models as critical, practical design guidelines for engineering mutual understanding—especially across the opacity of large language models or deep generative networks—are underdeveloped. O'neill *et al.* [14] and Wang *et al.* [24] gesture toward knowledge-extraction pipelines, but systematic methods for aligning AI and human representations of complex material spaces are still exploratory [30, 31].

Cultural and institutional barriers to AI as collaborator. Academic reward structures, publication norms, and laboratory hierarchies continue to privilege human-centric authorship and decision-making, discouraging the epistemic redistribution required for true partnership. Chauhan and Otyepka *et al.* [9] note that institutional inertia slows adoption of collaborative framings, while seed references such as Kim and Nam [1] and Mitchell [2] remind us that cultural metaphors shape technology uptake more powerfully than technical capability alone.

Evaluation metrics for collaboration quality (not just task performance). Current assessment practices focus almost exclusively on accuracy, speed, or discovery rate [9, 26], ignoring relational metrics such as mutual learning, creativity amplification, or long-term trust sustainability. The absence of such metrics, highlighted by Gonzalez *et al.* and Veitch *et al.* in broader HCI contexts, prevents the field from distinguishing productive collaboration from superficial tool use.

These gaps are mutually reinforcing: without empirical studies, validated models cannot emerge; without trust and shared understanding, research and cultural barriers persist. Addressing them demands a deliberate shift in research priorities away from purely algorithmic innovation toward socio-technical investigation of the interface itself.

Future Directions

To close the identified gaps and realize the full potential of AI as a collaborator, six targeted future research directions are proposed. Each direction builds on the dimensions, models, and shifts articulated throughout the review and offers concrete pathways for advancing both theory and practice in materials science.

Empirical studies of real AI-scientist collaboration. Longitudinal, ethnographic, and mixed-methods investigations embedded in active materials laboratories are urgently needed. Such studies should track autonomy calibration, trust trajectories, and role flexibility across complete discovery cycles, moving beyond controlled prototypes [7, 11, 13] to document lived collaboration in high-stakes projects.

Design guidelines for collaborative AI systems. Practical, dimension-informed guidelines must be developed that translate the six interface axes into actionable interface specifications. These guidelines should incorporate communication modalities from Jiang *et al.* [12] and shared-understanding techniques from Xu and Gao [27], ensuring future systems are engineered for partnership rather than tool-like subservience [15, 19, 22].

Training programs for human-AI teaming. Interdisciplinary curricula that prepare materials scientists to work effectively with autonomous and generative AI partners are essential. Training should address cognitive partnership risks [19], trust

repair strategies [24], and goal-alignment negotiation [32, 33], equipping researchers to engage AI as teammates rather than instruments [1, 2].

Metrics for collaboration quality and success. New evaluation frameworks must supplement traditional performance metrics with relational and epistemic indicators—mutual learning rates, creativity amplification scores, and trust sustainability indices. Such metrics, building on calls from Shi *et al.* and Gonzalez *et al.*, will enable rigorous comparison across models and dimensions.

Adaptive interfaces that evolve with the relationship. Interfaces capable of dynamically adjusting autonomy, communication richness, and role assignments as the human-AI relationship matures should be prototyped and tested. Drawing on the role flexibility observed in Liu *et al.* and Lu *et al.* [31], these adaptive systems would mirror the spectrum from tool to collaborator in real time [6, 35].

Ethical frameworks for AI collaborator roles. As AI assumes greater epistemic authority, explicit ethical frameworks are required to govern accountability, credit allocation, bias mitigation, and the boundaries of acceptable autonomy. These frameworks must address institutional barriers [8, 23] and ensure that collaboration enhances rather than displaces human scientific agency [16, 18].

Collectively, these directions chart a systematic research agenda that moves materials AI from incremental tool improvement to deliberate cultivation of collaborative capability. Implementation will require cross-disciplinary teams spanning HCI, AI ethics, and materials science, with the explicit goal of producing not only better algorithms but better partnerships.

Conclusion

This review has traced the conceptual evolution of the AI-materials scientist interface, documenting the persistent dominance of the tool metaphor, the five critical shifts that enable collaborative framings, the five existing interface models, and the six interdependent dimensions that define partnership quality. By synthesizing exactly 35 peer-reviewed publications from 2017–2025, the analysis reveals both the progress already achieved—evident in autonomous networks, shared cognitive systems, and multimodal feedback loops —and the substantial gaps that still separate current practice from genuine collaboration. The limitations of the tool metaphor, the scarcity of empirical teaming studies, and the absence of validated metrics for relational success all underscore that materials science stands at a pivotal transition.

The central thesis is clear: treating AI merely as a calculator, predictor, or database no longer suffices. As systems acquire agency, adaptivity, and generative capacity, the field must intentionally redesign the interface around autonomy, communication, trust, shared understanding, goal alignment, and role flexibility. Only through such redesign can the epistemic and creative potential of human-AI partnership be realized. Future work must therefore prioritize the six proposed directions, transforming conceptual models into deployable, trustworthy, and ethically grounded collaborative platforms.

Materials discovery has always been a deeply human endeavor. By embracing AI as collaborator rather than tool, the discipline can augment that humanity—accelerating insight, expanding imaginative reach, and deepening collective understanding of matter itself—while preserving the values of scientific rigor, creativity, and responsibility that define the field. The transition is not inevitable; it is a deliberate choice that the community must now make with clarity and foresight.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 12 Sep 2024 Revised: 24 Oct 2024 Accepted: 09 Dec 2024
Published online: 18 January 2025

References

Kim CM, Nam TJ. Exploration on everyday objects as an IoT control interface. In: Proceedings of the 2022 ACM Designing Interactive Systems Conference. New York, NY: ACM; 2022. p. 1654-68.

Mitchell M. Artificial intelligence: A guide for thinking humans. New York, NY: Farrar, Straus and Giroux; 2019.

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.

Häse F, Roch LM, Aspuru-Guzik A. Next-generation experimentation with self-driving laboratories. Trends Chem. 2019;1(3):282-91.

MacLeod BP, Parlane FG, Morrissey TD, Häse F, Roch LM, Dettelbach KE, et al. Self-driving laboratory for accelerated discovery of thin-film materials. Sci Adv. 2020;6(20):eaaz8867.

Stach E, DeCost B, Kusne AG, Hattrick-Simpers J, Brown KA, Reyes KG, et al. Autonomous experimentation systems for materials development: A community perspective. Matter. 2021;4(9):2702-26.

Otyepka M, Pykal M, Otyepka M. Advancing materials discovery through artificial intelligence. Appl Mater Today. 2025;47:102981.

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.

Kitano H. Nobel turing challenge: Creating the engine for scientific discovery. npj Syst Biol Appl. 2021;7(1):29.

Karpatne A, Deshwal A, Jia X, Ding W, Steinbach M, Zhang A, et al. AI-enabled scientific revolution in the age of generative AI: Second NSF workshop report. npj Artif Intell. 2025;1(1):18.

Shao E, Wang Y, Qian Y, Pan Z, Liu H, Wang D. SciSciGPT: Advancing human-AI collaboration in the science of science. Nat Comput Sci. 2025:1-5.

O'Neill T, McNeese N, Barron A, Schelble B. Human–autonomy teaming: A review and analysis of the empirical literature. Hum Factors. 2022;64(5):904-38.

O'Neill TA, Flathmann C, McNeese NJ, Salas E. Human-autonomy teaming: Need for a guiding team-based framework? Comput Hum Behav. 2023;146:107762.

McNeese NJ, Flathmann C, O'Neill TA, Salas E. Stepping out of the shadow of human-human teaming: Crafting a unique identity for human-autonomy teams. Comput Hum Behav. 2023;148:107874.

Andrews RW, Lilly JM, Srivastava D, Feigh KM. The role of shared mental models in human-AI teams: A theoretical review. Theor Issues Ergon Sci. 2023;24(2):129-75.

Bansal G, Nushi B, Kamar E, Lasecki WS, Weld DS, Horvitz E. Beyond accuracy: The role of mental models in human-AI team performance. In: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. Palo Alto, CA: AAAI Press; 2019;7(1):2-11.

Bansal G, Nushi B, Kamar E, Horvitz E, Weld DS. Is the most accurate AI the best teammate? Optimizing AI for teamwork. In: Proceedings of the AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press; 2021;35(13):11405-14.

Bansal G, Wu T, Zhou J, Fok R, Nushi B, Kamar E, et al. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. New York, NY: ACM; 2021. p. 1-16.

Hemmer P, Schemmer M, Kühl N, Vössing M, Satzger G. Complementarity in human-AI collaboration: Concept, sources, and evidence. Eur J Inf Syst. 2025;34(6):979-1002.

Senoner J, Schallmoser S, Kratzwald B, Feuerriegel S, Netland T. Explainable AI improves task performance in human–AI collaboration. Sci Rep. 2024;14(1):31150.

Wu S, Liu Y, Ruan M, Chen S, Xie XY. Human-generative AI collaboration enhances task performance but undermines human's intrinsic motivation. Sci Rep. 2025;15(1):15105.

Wang Z, Cao L, Jin Q, Chan J, Wan N, Afzali B, et al. A foundation model for human-AI collaboration in medical literature mining. Nat Commun. 2025;16(1):8361.

Zhang G, Chong L, Kotovsky K, Cagan J. Trust in an AI versus a human teammate: The effects of teammate identity and performance on human-AI cooperation. Comput Hum Behav. 2023;139:107536.

Flathmann C, Schelble BG, Rosopa PJ, McNeese NJ, Mallick R, Madathil KC. Examining the impact of varying levels of AI teammate influence on human-AI teams. Int J Hum Comput Stud. 2023;177:103061.

Xu W, Gao Z. Applying HCAI in developing effective human-AI teaming: A perspective from human-AI joint cognitive systems. Interactions. 2024;31(1):32-7.

Caldwell S, Sweetser P, O'donnell N, Knight MJ, Aitchison M, Gedeon T, et al. An agile new research framework for hybrid human-AI teaming: Trust, transparency, and transferability. ACM Trans Interact Intell Syst. 2022;12(3):1-36.

Zhang R, McNeese NJ, Freeman G, Musick G. An ideal human expectations of AI teammates in human-AI teaming. Proc ACM Hum Comput Interact. 2021;4(CSCW3):1-25.

Attig C, Wollstadt P, Schrills T, Franke T, Wiebel-Herboth CB. More than task performance: Developing new criteria for successful human-AI teaming using the cooperative card game hanabi. In: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. New York, NY: ACM; 2024. p. 1-11.

Duan W, Flathmann C, McNeese N, Scalia MJ, Zhang R, Gorman J, et al. Trusting autonomous teammates in human-AI teams-a literature review. In: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. New York, NY: ACM; 2025. p. 1-23.

Hu M, Zhang G, Chong L, Cagan J, Goucher-Lambert K. How being outvoted by AI teammates impacts human-AI collaboration. Int J Hum Comput Interact. 2025;41(7):4049-66.

Winter J. AI teammates and human performance: Evidence for commitment deficits. Comput Hum Behav Rep. 2025;20:100828.

Breckner K, Neumayr T, Streit M, Augstein M. Personalized complementarity in human-AI collaboration. Bonn: Gesellschaft für Informatik eV; 2024. p. 10-18420.

Rodani T. Reliable AI in material science: A fair-by-design path from data to services. Available from: https://ricerca.unityfvg.it/entities/publication/88217b2b-b426-41d5-ac14-2d970f8877fa/details