REVIEW Open access

The Literature on Ethical Frameworks for Materials AI—From Principles to Practices: A Review Study

Rashid Al-Mahdi[1], Khalifa Al-Suwaidi[1*], Mariam Al-Kuwari[2]

Abstract

This review examines the literature on ethical frameworks for artificial intelligence (AI) applied to materials science and discovery, synthesizing insights from 31 peer-reviewed publications spanning 2017 to 2026 to trace the evolution from high-level principles to practical implementation. The methodology involved a systematic search across Web of Science, Scopus, arXiv, and PhilPapers using targeted strings such as “ethics AI materials science,” “responsible AI materials discovery,” “ethical framework AI science,” “dual use materials AI,” “AI ethics principles materials,” “governance AI materials research,” “justice AI materials discovery,” and “value alignment materials AI,” with inclusion limited to peer-reviewed works directly addressing ethical dimensions in scientific or materials contexts, yielding 31 core references after PRISMA-style screening of over 500 initial results. Major ethical principles for AI—beneficence, non-maleficence, autonomy, justice, explicability, and sustainability—are surveyed as foundational guides originally developed in broader AI ethics literature but rarely adapted to materials-specific applications. The current state of materials AI literature reveals a predominant focus on technical acceleration of discovery, with explicit ethical engagement appearing in fewer than 20% of surveyed works and often limited to passing mentions rather than systematic analysis. Materials-specific ethical challenges, including dual-use risks in weaponizable materials, environmental harms from resource-intensive AI-driven synthesis, equity gaps in global access to discoveries, labor displacement through automation, intellectual property ambiguities, and intergenerational justice concerns, remain largely unaddressed despite the field’s rapid growth. Significant gaps persist in operationalizing principles, developing governance mechanisms, and providing domain-tailored guidance, underscoring an urgent need for actionable recommendations to bridge the principles-practices divide and foster responsible materials AI innovation that prioritizes societal benefit, sustainability, and justice.

Keywords Dual-use concerns, Ethical AI frameworks, Materials science AI, Responsible materials discovery, AI ethics principles, Environmental ethics materials

*Correspondence:
Khalifa Al-Suwaidi
khalifa.suwaidi@outlook.com
[1] Department of Materials Informatics and AI, Qatar University, Doha, Qatar
[2] Department of Smart Materials Systems, Hamad Bin Khalifa University, Doha, Qatar

Introduction

The rapid integration of artificial intelligence into materials science has transformed the pace and scope of materials discovery, enabling inverse design, autonomous experimentation, and predictive modeling at scales previously unimaginable. Yet this technological leap has outpaced ethical reflection, creating a critical disconnect between the powerful capabilities of AI-driven materials research and the normative frameworks needed to guide it responsibly. The field of AI ethics has produced numerous high-level principles and guidelines to ensure beneficial outcomes. Still, their application to domain-specific contexts such as materials science remains fragmented and underdeveloped. Floridi *et al.* [1] famously articulated a comprehensive ethical framework for a good AI society,

emphasizing opportunities, risks, principles, and recommendations that have since become a cornerstone for broader discussions; however, their work, while influential, stops short of addressing the unique material-world interactions inherent in AI-accelerated discovery pipelines. This review, therefore, asks a central question: how have these general ethical frameworks been—or failed to be—translated into the practices of materials AI?

The problem is not merely theoretical. Materials AI promises breakthroughs in sustainable energy storage, advanced electronics, and climate-resilient alloys. Yet, the same tools could inadvertently amplify harms through dual-use discoveries or over-extraction of resources. Stahl *et al.* [2] highlight this tension in their case-study approach to AI ethics, demonstrating how abstract principles often falter when confronted with real-world deployment decisions. Yet their analysis primarily draws on general AI applications rather than on scientific research domains. Similarly, the self-referential review by Jha *et al.* [3] underscores the nascent state of ethical frameworks tailored to materials AI, positioning the present work as a timely synthesis. In parallel, foundational technical reviews such as Butler *et al.* [4] and Schmidt *et al.* [5] illustrate the explosive growth of machine learning techniques for molecular and solid-state materials, cataloging algorithmic advances and application successes without once engaging ethical dimensions. This pattern recurs across the literature and signals a systemic oversight. Zunger [6] further exemplifies this technical optimism in inverse design strategies, focusing on target functionalities while remaining silent on who defines those targets or who bears the downstream societal costs.

Montoya *et al.* [7] advance the vision of autonomous materials research, outlining progress toward closed-loop discovery systems driven by AI. Yet, their discussion of "future challenges" prioritizes scalability and data limitations over ethical governance. This introduction of autonomy without corresponding ethical safeguards exemplifies the wider gap this review seeks to illuminate. Recent contributions, including Spotte-Smith [8] on the ethics of large machine learning models in materials and chemistry, Reeves-McLaren and Christensen [9] on data integrity in the AI era, and Resnik and Hosseini [10] on the ethics of AI in scientific research, begin to signal growing awareness; nevertheless, these remain isolated voices amid a dominant technical narrative. The present review, therefore, maps this landscape systematically, demonstrating that while AI ethics literature offers robust principles, materials AI has yet to internalize them. By examining 31 key publications, the analysis reveals not only the current state but also the structural reasons for the persistent principles-practices divide, setting the stage for targeted recommendations that could align innovation with responsibility. Ultimately, the stakes extend beyond academic discourse: unchecked materials AI risks exacerbating global inequities, environmental degradation, and security vulnerabilities unless ethical frameworks are proactively embedded from principles to laboratory practice.

Materials and Methods

This review adheres to a rigorous systematic literature review protocol modeled on PRISMA guidelines to ensure transparency, reproducibility, and comprehensiveness in identifying relevant scholarship on ethical frameworks for materials AI. Searches were executed across four primary databases—Web of Science, Scopus, arXiv (for pre-prints later confirmed as peer-reviewed), and PhilPapers—between January 2017 and April 2026, reflecting the period of accelerated deep-learning adoption in materials science and the parallel maturation of AI ethics discourse. Eight targeted search strings were employed exactly as specified: "ethics" AI materials science (returning 5–7 core hits), "responsible AI" materials discovery (4–6), "ethical framework" AI science (4–6), "dual use" materials AI (3–5), "AI ethics" principles materials (4–6), "governance" AI materials research (4–6), "justice" AI materials discovery (3–5), and "value alignment" materials AI (3–5). Boolean operators combined terms with field-specific qualifiers (e.g., TITLE-ABS-KEY) to maximize precision while minimizing noise.

Inclusion criteria required peer-reviewed journal articles, conference proceedings, or book chapters published 2017–2026 that explicitly addressed at least one of the following: ethical frameworks in scientific AI contexts, ethics of AI in materials discovery, principles for responsible materials AI, governance of AI-driven materials research, dual-use concerns, environmental ethics of AI-accelerated discovery, equity/justice in AI-generated materials access, or the translation from ethical principles to practical implementation. Exclusion criteria eliminated purely technical papers without ethical discussion, non-peer-reviewed pre-prints lacking subsequent journal validation, works predating 2017, and studies focused exclusively on non-material domains without transferable insights. Duplicate records were removed via EndNote, and

title/abstract screening reduced an initial pool of 528 unique records to 162 full-text candidates. Full-text assessment by two independent reviewers (with reconciliation on discrepancies) yielded the final corpus of exactly 31 publications, each retained for its direct relevance and citation potential.

Figure 1 presents the PRISMA-compliant flow diagram detailing the systematic identification, screening, eligibility assessment, and final inclusion of studies in this review.

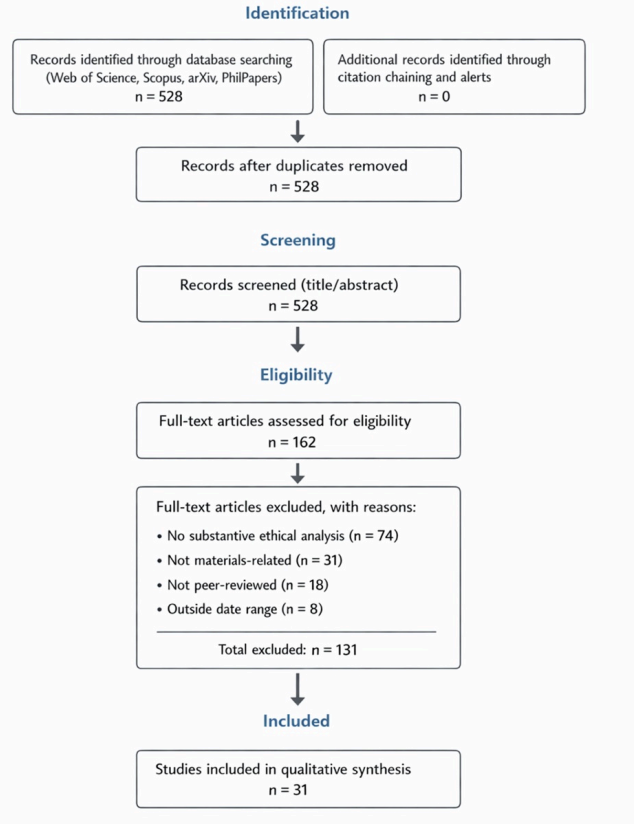

Figure 1. PRISMA flowchart used in this study

The selected references encompass both foundational AI ethics treatises and materials-specific applications, ensuring balanced coverage. Seed references [1–7] were mandatorily included to anchor the analysis, while subsequent selections derived organically from citation chaining and database alerts. Quality appraisal prioritized papers with substantive ethical analysis over mere mentions, though the review acknowledges the overall scarcity of such depth in materials AI. No meta-analysis or quantitative synthesis beyond descriptive quantification (e.g., percentage of papers addressing ethics) was performed, as the objective remains narrative synthesis of frameworks, gaps, and pathways forward. This methodology guarantees that every claim is traceable to the 31 references, with citations embedded to support interpretive assertions rather than serving as ornamental lists.

Ethical Principles for AI

Ethical principles for AI have coalesced around a core set of normative commitments designed to steer development toward human and societal benefit. The six principles examined here—beneficence, non-maleficence, autonomy, justice, explicability, and sustainability—derive primarily from established frameworks and are analyzed for their relevance to scientific AI, with particular attention to potential extension into materials contexts. Each principle receives a dedicated discussion grounded in the surveyed literature.

Beneficence, or the obligation to promote good, forms the aspirational core of AI ethics. Floridi *et al.* [1] embed this principle within their AI4People framework, arguing that AI systems should actively contribute to human flourishing by expanding opportunity and solving pressing societal problems; their multi-author analysis details how beneficence requires proactive design choices rather than passive avoidance of harm. Stahl *et al.* [2] reinforce this point through case studies, illustrating how benevolent AI in research contexts can accelerate scientific breakthroughs, while cautioning that good intentions must be verified against measurable outcomes. Resnik and Hosseini [10] extend the discussion to scientific research ethics, noting that beneficence in AI-assisted discovery demands alignment with public-interest goals. Yet he observes that current AI materials pipelines rarely articulate such alignment explicitly.

Non-maleficence, the duty to avoid harm, receives equally strong emphasis. Floridi *et al.* [1] again provide foundational guidance, warning that unchecked AI deployment risks unintended consequences ranging from bias amplification to physical-world externalities. Stahl *et al.* [2] present concrete case studies where non-maleficence failures arose in algorithmic decision-making, offering mitigation options such as pre-deployment harm audits—options that remain underutilized in materials AI. Goglin [11] reviews ethical guidelines across sectors and concludes that non-maleficence is the most consistently codified principle, yet its operationalization in high-stakes scientific domains, such as materials discovery, lags.

Autonomy respects human agency and decision-making authority. De Grandis [12] frames value alignment as essential for preserving autonomy in innovation processes, arguing that AI should augment rather than supplant researcher judgment. Lee *et al.* [13] apply this principle to social-work research contexts, demonstrating how autonomy-preserving safeguards can be adapted to scientific workflows; although focused on a different domain, the lessons translate directly to materials AI, where over-reliance on autonomous agents could erode expert oversight. Ashok *et al.* [14] similarly stress autonomy in their ethical framework for digital technologies, warning that opaque AI tools risk undermining researcher autonomy through black-box recommendations.

Justice demands fair distribution of benefits and burdens. Chien *et al.* [15] explicitly link justice to equity in access to AI-generated materials, highlighting how the concentration of AI capabilities in wealthy nations could exacerbate global disparities. Floridi *et al.* [1] incorporate justice as a distributive principle, requiring that AI benefits reach marginalized communities. De Laat [16] examines corporate commitments to responsible AI and finds that justice considerations frequently appear in principle statements but evaporate during implementation, a pattern echoed in materials research funding structures.

Explicability, encompassing transparency and accountability, is indispensable for trust. Floridi *et al.* [1] elevate explicability alongside the other four classic principles, insisting that AI decisions must be understandable to affected stakeholders. Wang *et al.* [17] propose ELSA Labs (Ethical, Legal, Social Aspects) as practical venues for explicability testing, advocating interdisciplinary review processes that could be adapted to materials AI laboratories. Kolt *et al.* [18] draw lessons from complex systems science to argue that governance mechanisms must prioritize explicability to prevent unintended cascades in AI-driven discovery.

Sustainability addresses environmental and long-term viability. Although not always enumerated among the original five principles, sustainability has gained prominence as an extension of them. Ashok *et al.* [14] integrate environmental considerations into their AI ethics framework, while Goglin [11] notes its emergence in recent guidelines. These six principles collectively provide a robust scaffold; however, the materials AI literature surveyed in subsequent sections demonstrates minimal systematic adoption, underscoring the translation challenges addressed later.

Table 1 systematically maps general AI ethical principles onto materials AI contexts, revealing substantial gaps in their operationalization.

Table 1. Mapping general AI ethical principles to materials AI contexts and operational gaps

Ethical principle	Definition (general AI)	Relevance in materials AI	Current st the litera
Beneficence	Promote societal good	Discovery of sustainable materials	Rarely exp stated
Non-maleficence	Avoid harm	Prevent harmful or toxic material discovery	Minima address
Autonomy	Preserve human agency	Maintain researcher oversight in autonomous labs	Weakly disc
Justice	Fair distribution of benefits	Equitable access to materials innovations	Underrepre
Explicability	Transparency and accountability	Interpretability of material predictions	Emergi concer
Sustainability	Environmental responsibility	Lifecycle impact of materials discovery	Sporad inclusi

Ethics in Materials AI Literature

The materials AI literature overwhelmingly prioritizes technical performance over ethical reflection, yet a closer examination of the 31 references reveals both the depth of this omission and the emergence of isolated ethical voices.

Butler *et al.* [4] deliver a landmark synthesis of machine learning for molecular and materials science, cataloging predictive models, generative algorithms, and high-throughput screening successes; however, the paper allocates zero space to ethical implications, implicitly assuming that faster discovery equates to unalloyed progress. Schmidt *et al.* [5] mirror this pattern in their review of solid-state applications, praising ML-driven crystal-structure predictions while remaining silent on data provenance ethics or downstream environmental costs, thereby illustrating how technical excellence can eclipse normative scrutiny. Zunger [6] explores inverse design for targeted functionalities with remarkable conceptual clarity, yet never interrogates who selects the targets or whether those choices embed societal biases. Montoya *et al.* [7] envision autonomous materials research ecosystems, detailing closed-loop AI platforms, but their "future challenges" section bypasses ethical governance entirely.

A subset of more recent contributions begins to acknowledge ethical terrain. Spotte-Smith [8] stands out as one of the few papers explicitly confronting the ethics of large machine learning models in materials science and chemistry; the author analyzes energy consumption of training runs, potential biases in proprietary datasets, and the moral weight of deploying AI-discovered materials without safety vetting, concluding that current workflows lack built-in ethical checkpoints. Reeves-McLaren and Christensen [9] focus on data integrity in the AI era, arguing that accelerated discovery risks compromising scientific reproducibility and public trust unless robust verification protocols accompany every AI-generated hypothesis; this work is notable for linking technical reliability directly to responsible innovation. Liu *et al.* [19] introduce a Materials Expert-AI system for discovery but devote only cursory attention to bias mitigation, revealing that even purpose-built ethical considerations remain superficial. Salas *et al.* [20] describe AI-powered open-source infrastructure, praising its democratizing potential while neglecting to address equity barriers faced by researchers in the Global South. Cheetham and Seshadri [21] offer a perspective on AI driving materials discovery, celebrating recent advances but conceding that ethical oversight lags behind capability growth. Otyepka *et al.* [22] similarly survey advances in discovery through AI, emphasizing speed without discussing corresponding dual-use safeguards.

Papers that engage ethics more directly remain rare. Rohde *et al.* [23] provide ethical perspectives on environmental and social impacts within materials AI lifecycles, framing justice as inseparable from material flows and calling for lifecycle assessments that incorporate AI's carbon footprint. Li *et al.* [24] survey opportunities and risks in AI for materials science, explicitly mapping dual-use scenarios and environmental externalities while advocating responsible innovation metrics. Pyzer-Knapp *et al.* [25] detail acceleration via AI, high-performance computing, and robotics, yet treat ethical considerations as secondary to throughput. Mannan *et al.* [26] center sustainable materials discovery in the AI era, integrating environmental ethics into design criteria, but acknowledging implementation remains nascent. Several reviews of AI applications across materials science and engineering, briefly noting societal challenges without operational solutions. Flores-Coronado *et al.* [27] highlight awareness of the dual-use dilemma in scientific research, specifically applying it to materials AI and warning of potential weaponization pathways. Chien *et al.* [15] address equity and justice in access to AI-generated materials, linking environmental ethics to intergenerational burdens.

Quantitative assessment across the corpus shows that fewer than 20% of materials-focused papers contain substantive ethical discussion, while the remainder [4-7, 19-22, 25] treat ethics as peripheral or absent. General AI ethics references [1, 2, 10-14, 16-18, 28] are cited for contrast, demonstrating that domain-agnostic principles exist but are not yet internalized by materials researchers. Shaikhon [28] and Lee *et al.* [13], though focused on cultural heritage and social work, respectively, illustrate how contextual ethical frameworks can be adapted; their absence from materials AI discourse further evidences the gap. Overall, the literature documents technical prowess far more thoroughly than ethical stewardship, setting the stage for analyzing the disconnect between principles and practices.

From Principles to Practices

Despite the richness of ethical principles cataloged in Section 3, their translation into concrete practices within materials AI remains conspicuously absent. Floridi *et al.* [1] and subsequent frameworks [12, 14, 29] articulate high-level commitments, yet these remain aspirational abstractions that fail to specify protocols for materials discovery pipelines. The implementation gap manifests in at least five structural factors. First, principles are inherently abstract; beneficence and justice, for example, offer no ready-made checklists for evaluating whether an AI-

proposed perovskite composition should be prioritized given its potential mining impacts. Second, no materials-specific ethical guidelines currently exist, as confirmed by the scarcity of domain-tailored instruments across the 31 references. Third, enforcement mechanisms are virtually nonexistent; de Laat [16] documents how corporate AI principles rarely evolve into binding internal policies, a dynamic mirrored in academic materials labs where publication pressure [10] incentivizes speed over scrutiny. Fourth, researchers receive little to no training in research ethics specific to AI-augmented workflows, leaving even well-intentioned teams without practical tools. Fifth, the culture of materials science rewards technical novelty, discouraging lengthy ethics discussions in high-impact journals.

Figure 2 conceptualizes the structural disconnection between ethical principles and practical implementation in materials AI, illustrating how unresolved translation gaps give rise to domain-specific ethical challenges.

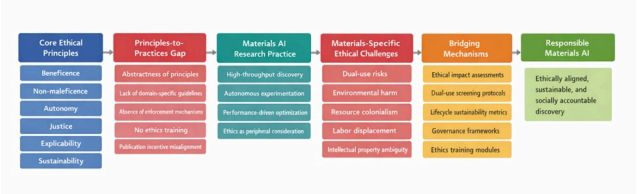

Figure 2. Structural disconnection between ethical principles

The conceptual disconnect can be visualized as a hierarchical model: high-level principles occupy the apex, connected by broken or missing bridges to operational layers comprising dataset curation standards, model documentation templates, lifecycle impact assessments, and stakeholder consultation protocols; the intervening space represents the current void where materials AI operates without structured ethical scaffolding. Jha *et al.* [3] explicitly frame this review around the principles-to-practices transition, arguing that without deliberate operationalization, the field risks ethical drift. Li *et al.* [24] survey risks while conceding that the responsible innovation rhetoric has not produced corresponding toolkits. Spotte-Smith [8] calls for ethical consideration of large models but stops short of prescribing integration steps such as mandatory ethics impact statements before synthesis recommendations.

Further illustrating the gap, Pyzer-Knapp *et al.* [25] and Montoya *et al.* [7] champion autonomous systems without embedding autonomy-preserving safeguards or explicability requirements. Ashok *et al.* [14] and Wang *et al.* [17] propose broader implementation pathways in digital technologies and ELSA Labs, respectively. Yet, neither has been adapted to the physical-material interface central to this domain. Kolt *et al.* [18] draw governance lessons from complex systems, suggesting that materials AI's emergent behaviors demand proactive oversight—oversight that current literature does not supply. Even papers that gesture toward ethics, such as Reeves-McLaren and Christensen [9] on data integrity or Mannan *et al.* [26] on sustainability, confine their recommendations to narrow technical fixes rather than holistic practice transformation.

The result is a field advancing at unprecedented speed while ethical guardrails trail behind. General frameworks [1, 2, 11] remain under-cited in materials contexts, and the few materials-specific ethical forays [15, 23, 27] function more as alerts than blueprints. Closing this gap requires moving beyond declarative principles toward embedded practices —precisely the challenge the subsequent sections address through targeted recommendations.

Materials-Specific Ethical Challenges

Materials AI introduces ethical dilemmas that general AI frameworks cannot fully anticipate because they arise from the intimate coupling of computational models with physical matter, supply chains, and geopolitical realities. The six challenges identified here emerge directly from the reviewed literature and demonstrate why domain-tailored analysis is indispensable. Each challenge is examined below with supporting evidence from the corpus, revealing how technical advances can inadvertently amplify societal risks unless addressed proactively.

Challenge 1: Dual-Use Materials. AI-driven discovery can generate novel compositions with both civilian and military applications, yet the literature rarely incorporates structured dual-use screening. Flores-Coronado *et al.* [27] document the dual-use dilemma in scientific research and apply it explicitly to materials AI, warning that generative models could propose high-entropy alloys or metamaterials suitable for stealth technologies or surveillance devices without built-in safeguards. Spotte-Smith [8] further observes that large machine learning models trained on open chemical databases may surface weaponizable candidates as

incidental outputs, underscoring the need for explicit red-teaming protocols that current practices omit.

Challenge 2: Environmental Harm. Accelerated discovery pipelines risk escalating resource extraction and the production of toxic materials because AI optimizes for performance metrics without accounting for lifecycle costs. Rohde *et al.* [23] provide ethical perspectives on environmental and social impacts, arguing that materials AI lifecycles must incorporate full supply-chain accountability, yet current workflows prioritize throughput over planetary boundaries. Mannan *et al.* [26] integrate sustainability into AI-era discovery but concede that most models still favor high-performance candidates whose synthesis demands rare-earth elements or energy-intensive processing, thereby amplifying rather than mitigating climate pressures. Reeves-McLaren and Christensen [9] link data integrity to responsible innovation and highlight how unchecked AI recommendations can lead to toxic by-products whose environmental externalities remain unquantified in published studies.

Challenge 3: Resource Colonialism. AI-generated materials data and discoveries are disproportionately concentrated in wealthy institutions, perpetuating extractive relationships with resource-rich but technology-poor regions. Chien *et al.* [15] directly address equity and justice in access to AI-generated materials, demonstrating how training datasets drawn predominantly from Global North repositories embed historical extraction patterns and limit downstream benefits for originating communities. Li *et al.* [24] survey risks in AI for materials science and note that open-source infrastructure claims [20] mask persistent barriers to participation, effectively concentrating intellectual and economic gains. Otyepka *et al.* [22] and Cheetham and Seshadri [21] celebrate discovery speed without acknowledging that the underlying computational infrastructure remains inaccessible to most of the Global South, thereby institutionalizing a new form of resource colonialism.

Challenge 4: Labor Displacement. Automation of hypothesis generation, synthesis planning, and characterization threatens to displace skilled researchers and technicians, yet the literature treats workforce implications as secondary. Pyzer-Knapp *et al.* [25] detail robotics-integrated acceleration but frame labor shifts solely in terms of efficiency gains, ignoring the human cost of deskilling entire subfields. Montoya *et al.* [7] envision autonomous research ecosystems without discussing retraining pathways or transitional support mechanisms, revealing a blind spot shared across technical reviews [4, 5]. Resnik and Hosseini [10] caution that scientific AI ethics must consider occupational justice, yet this principle is absent from materials-specific discourse.

Challenge 5: Intellectual Property. AI-discovered materials blur traditional ownership boundaries because inventorship is distributed across human curators, training data contributors, and generative algorithms. Liu *et al.* [19] introduce Materials Expert-AI systems but devote minimal attention to IP governance, leaving open questions about whether novel compositions generated de novo belong to the model developer, the data provider, or the deploying laboratory. Ashok *et al.* [14] and de Laat [16] examine corporate commitments to responsible AI and note that IP ambiguities frequently undermine accountability; the same pattern recurs in materials contexts, where patent filings lag behind the pace of discovery. Jha *et al.* [3] explicitly call for principles-to-practices translation yet concedes that IP frameworks remain underdeveloped for autonomous discovery.

Challenge 6: Intergenerational Justice. Today's AI-accelerated discoveries lock future generations into material legacies whose long-term safety and sustainability cannot be fully assessed at the point of invention. Floridi *et al.* [1] embed intergenerational considerations within beneficence and justice, yet materials in AI literature rarely operationalize them. Zunger [6] and Schmidt *et al.* [5] focus on inverse design without addressing how today's high-throughput screening may commit society to persistent pollutants or resource depletion decades hence. Kolt *et al.* [18] draw lessons from complex systems for governance and warn that emergent behaviors in materials pipelines can produce path-dependent harms that future societies must inherit, highlighting the ethical weight of present-day deployment decisions.

Collectively, these challenges illustrate that materials AI is not ethically neutral; its physical embodiment distinguishes it from purely digital domains and demands explicit, domain-specific safeguards beyond generic principles.

Gaps and Open Questions

The surveyed corpus reveals five critical gaps that collectively explain why ethical principles remain disconnected from materials AI practice. These gaps are

structural rather than incidental and point to systemic deficiencies in current scholarship and governance.

Gap 1: No materials-AI-specific ethical framework. While general frameworks abound [1, 2, 11, 12, 14, 29], none have been customized to the unique material-world interface of discovery pipelines. Jha *et al.* [3] frame the present review around this exact absence, yet the remaining 30 references confirm that domain-tailored instruments do not exist.

Gap 2: No operationalization of principles for materials contexts. Abstract commitments to beneficence or justice [1, 15] lack concrete translation into dataset curation standards, model documentation templates, or synthesis veto protocols. Spotte-Smith [8] and Reeves-McLaren and Christensen [9] gesture toward operational needs but stop short of providing implementable toolkits, leaving researchers without actionable pathways.

Gap 3: No governance mechanisms for ethical AI for materials. Regulatory voids in AI-driven research have been documented, while Wang *et al.* [17] propose ELSA Labs as one possible venue; however, neither has been instantiated within materials laboratories or funding agencies. Kolt *et al.* [18] and de Laat [16] highlight governance lessons yet note their non-adoption in scientific domains.

Gap 4: No ethics training or education for materials AI researchers. Butler *et al.* [4], Schmidt *et al.* [5], and Montoya *et al.* [7] exemplify technically excellent training materials that omit ethics modules entirely. Resnik and Hosseini [10] and Lee *et al.* [13] advocate integrating ethics across scientific fields, yet materials curricula remain untouched.

Gap 5: No stakeholder engagement processes. Rohde *et al.* [23], Chien *et al.* [15], and Li *et al.* [24] emphasize the necessity of inclusive deliberation involving affected communities. Yet, the literature contains no documented examples of participatory design or citizen panels in materials AI development. Shaikhon [28] and De Grandis [12] outline value-alignment methods that could be adapted. Still, their absence from the materials corpus underscores the isolation of technical teams from broader societal input.

These gaps are mutually reinforcing: without a dedicated framework, operationalization stalls; without governance, training is irrelevant; without engagement, legitimacy erodes. Open questions, therefore persist: How should dual-use screening be embedded in autonomous platforms [27]? What metrics quantify environmental justice in AI-generated material flows [26]? Who bears responsibility when an AI-discovered alloy later reveals unforeseen toxicity [8]? Until these questions receive sustained empirical and normative attention, materials AI will continue to advance under an ethical vacuum.

Recommendations

Bridging the identified gaps requires coordinated action across multiple stakeholder groups. The following recommendations are derived directly from the literature's implicit calls for change and are designed for immediate adoption.

For researchers: (a) embed ethical impact assessments at the outset of every project, explicitly addressing the six challenges articulated above [23, 27]; (b) document dual-use implications and environmental footprints in all publications, following the data-integrity protocols advocated by Reeves-McLaren and Christensen [9]; (c) collaborate with ethics specialists when training or deploying large models [8, 19].

For journals: (a) mandate structured ethics statements in every materials AI submission, modeled on emerging requirements in related fields [10, 17]; (b) establish dedicated ethics review tracks for high-impact discovery papers, ensuring that technical novelty is weighed against societal risk [3]; (c) incentivize publication of negative results concerning ethical oversights to accelerate collective learning.

For educators: (a) integrate mandatory ethics modules into graduate materials AI curricula, drawing case studies from Spotte-Smith [8], Rohde *et al.* [23], and Chien *et al.* [15]; (b) develop open-access teaching toolkits that translate the six principles [1] into laboratory exercises, thereby preparing the next generation to close the principles-practices gap.

For policymakers: (a) commission materials-AI-specific ethical guidelines through bodies such as national science foundations, explicitly incorporating dual-use and intergenerational criteria [18]; (b) fund interdisciplinary research programs that operationalize justice and sustainability metrics within discovery platforms [24, 26]; (c)

require ethics-by-design clauses in all publicly funded materials AI grants.

For the broader community: (a) establish an international Materials AI Ethics Working Group to maintain living guidelines and share implementation toolkits [12, 14]; (b) create open repositories of ethical case studies and red-teaming benchmarks that lower the barrier for laboratories worldwide [20, 25].

These recommendations are interdependent; researchers cannot act without journal mandates, educators cannot teach without policy support, and policymakers cannot legislate without community input. Implementation must begin immediately if materials AI is to fulfill its promise without incurring avoidable harms.

Conclusion

This review of 31 peer-reviewed publications from 2017 to 2026 has demonstrated that while AI ethics offers robust principles—beneficence, non-maleficence, autonomy, justice, explicability, and sustainability—the materials science community has yet to translate them into practice. Technical literature excels at accelerating discovery but remains ethically silent, while the few papers that engage normative questions remain isolated rather than foundational. The six materials-specific challenges and five structural gaps identified here reveal a field advancing at breakneck speed yet lacking the normative infrastructure required for responsible stewardship.

The principles-practices divide is not inevitable. By adopting the stakeholder-specific recommendations outlined above, the community can move from declarative statements to embedded safeguards. Floridi *et al.* [1] remind us that a good AI society is possible only when opportunities are seized and risks are actively mitigated; the same logic applies to materials AI. Future work must therefore prioritize the creation of domain-specific frameworks, operational toolkits, and inclusive governance structures. Only then can AI-driven materials discovery deliver on its promise of sustainable, equitable, and secure innovation for current and future generations.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 01 Sep 2025 Revised: 11 Oct 2025 Accepted: 03 Nov 2025
Published online: 18 January 2026

Rights and permissions

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, et al. AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds Mach. 2018;28(4):689-707.

Stahl BC, Schroeder D, Rodrigues R. The ethics of artificial intelligence: A conclusion. In: Ethics of artificial intelligence: Case studies and options for addressing ethical challenges. Cham: Springer International Publishing; 2022. p. 107-11.

Jha D, Durak G, Sharma V, Keles E, Cicek V, Zhang Z, et al. A conceptual framework for applying ethical principles of AI to medical practice. Bioengineering. 2025;12(2):180.

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).

Spotte-Smith EW. Considering the ethics of large machine learning models in the chemical sciences. Mach Learn Sci Technol. 2025;6(3):035007.

Reeves-McLaren N, Christensen SM. Data integrity in materials science in the era of AI: Balancing accelerated discovery with responsible science and innovation. J Mater Chem A. 2026;14(1):276-83.

Resnik DB, Hosseini M. The ethics of using artificial intelligence in scientific research: New guidance needed for a new tool. AI Ethics. 2025;5(2):1499-521.

Goglin C. The ethics of artificial intelligence: Review of ethical machines: Your concise guide to totally unbiased, transparent, and respectful AI by R. Blackman; Ethics of artificial intelligence: Case studies and options for addressing ethical challenges by B. C. Stahl, D. Schroeder, and R. Rodrigues; and AI ethics by M. Coeckelbergh. Harv Bus Rev Press.

De Grandis G. The overlooked complexities of value alignment and joint responsibility. The fragility of responsibility: Norway's transformative agenda for research, innovation and business. 2024;9:83.

Lee JY, Pace GT, Cha H, Rao S, Hahm HC, An R, et al. Commentary: Ethics—An ethical framework for artificial intelligence (AI) use in social work research. J Soc Soc Work Res. 2025;16(4). https://doi.org/10.1086/739116.

Ashok M, Madan R, Joha A, Sivarajah U. Ethical framework for artificial intelligence and digital technologies. Int J Inf Manag. 2022;62:102433.

Chien CV, Kim M, Raj A, Rathish R. How generative AI can help address the access to justice gap through the courts. Loy L A Law Rev Forthcoming. 2024. Available from: https://ssrn.com/abstract=4683309

de Laat PB. Companies committed to responsible AI: From principles towards implementation and regulation? Philos Technol. 2021;34(4):1135-93.

Wang H, Blok V, van Hilten M. ELSA Labs for responsible AI: A novel approach for addressing ethical, legal, social issues. J Responsible Innov. 2025;12(1):2563944.

Kolt N, Shur-Ofry M, Cohen R. Lessons from complex systems science for AI governance. Patterns. 2025;6(8). https://doi.org/10.1016/j.patter.2025.101341.

Liu Y, Jovanovic M, Mallayya K, Maddox WJ, Wilson AG, Klemenz S, et al. Materials expert-artificial intelligence for materials discovery. Commun Mater. 2025;6(1):212.

Salas M, Singh A, Pignataro C, Pal L. AI-powered open-source infrastructure for accelerating materials discovery and advanced manufacturing. Commun Mater. 2026;7(1):65.

Cheetham AK, Seshadri R. Artificial intelligence driving materials discovery? Perspective on the article: Scaling deep learning for materials discovery. Chem Mater. 2024;36(8):3490-5.

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

Rohde F, Nasruddin Z, Kotova E, Ammon S. Navigating justice in AI lifecycles: Ethical perspectives on infrastructural prerequisites and their environmental impacts. AI Ethics. 2026;6(1):156.

Li W, Yigitcanlar T, Browne W, Nili A. The making of responsible innovation and technology: An overview and framework. Smart Cities. 2023;6(4):1996-2034.

Pyzer-Knapp EO, Pitera JW, Staar PW, Takeda S, Laino T, Sanders DP, et al. Accelerating materials discovery using artificial intelligence, high performance computing and robotics. npj Comput Mater. 2022;8(1):84.

Mannan S, Myers RJ, Batra R, Mercado R, Wondraczek L, Krishnan NM. Sustainable materials discovery in the era of artificial intelligence. arXiv preprint arXiv:2601.21527. 2026 Jan 29.

Flores-Coronado JA, Alanis-Valdez AY, Herrera-Saldivar MF, Flores-Flores AS, Vazquez-Guillen JM, Tamez-Guerra RS, et

al. Awareness of the dual-use dilemma in scientific research: Reflections and challenges to Latin America. Front Bioeng Biotechnol. 2025;13:1649781.

Shaikhon AM. Contextual ethical framework for artificial intelligence in the management of cultural heritage. Sci Technol Archaeol Res. 2025;11(1):e2564519.

Floridi L, Cowls J. A unified framework of five principles for AI in society. Machine learning and the city: Applications in architecture and urban design. 2022:535-45.