ORIGINAL RESEARCH **Open access**

A Conceptual Typology of Scientific Harm from Misused Materials AI

Chinedu Okafor[1*], Amina Bello[1]

Abstract

In the rapidly expanding field of materials artificial intelligence, the term “harm” appears with increasing frequency yet remains strikingly ambiguous, applied interchangeably to environmental degradation caused by AI-accelerated discovery of resource-intensive compounds, to health risks arising from the deployment of novel toxic materials, to security threats posed by dual-use generative models, and even to epistemic distortions that undermine the reliability of scientific knowledge itself. This conceptual boundary paper argues that such vagueness is not merely semantic. Still, it fundamentally impedes responsible innovation in materials science, where AI systems now routinely propose molecular structures, optimize synthesis pathways, and guide autonomous experimentation. The present work therefore proposes a precise definition of scientific harm as negative consequences—direct, indirect, potential, or actual—arising specifically from the development, deployment, or misuse of AI systems in materials science that affect humans, environments, knowledge systems, or social structures in ways that are both causally traceable to the AI intervention and normatively undesirable within scientific practice. Building on this foundation, the paper advances a six-fold typology of scientific harm tailored to materials AI: environmental, health, security, epistemic, social, and economic. Each type is defined, mechanistically elaborated, illustrated with materials-specific examples drawn from the literature, and accompanied by detectable signatures that practitioners can monitor. The typology is further sharpened through explicit distinctions from nearby concepts such as risk, danger, misuse, unintended consequence, and side effect, thereby clarifying the unique normative force carried by the language of harm. Ultimately, the framework carries direct implications for harm-aware practice: authors, reviewers, and the broader materials AI community must move beyond vague risk disclaimers toward typed, evidence-based harm assessments that enable targeted mitigation and more ethically robust research trajectories. By furnishing these conceptual tools, the paper seeks to transform “harm” from an overloaded rhetorical placeholder into a precise analytic category capable of guiding the responsible maturation of materials artificial intelligence.

Keywords Materials AI, Scientific harm, Dual-use concerns, Epistemic harm, Environmental harm, Typology of harm

*Correspondence:
Chinedu Okafor
chinedu.okafor@gmail.com
[1] Department of Computational Materials Systems, University of Lagos, Lagos, Nigeria

Introduction

Materials artificial intelligence has transformed the pace and scope of discovery, enabling inverse design of novel compounds, autonomous laboratory workflows, and predictive modeling of properties once considered intractable. Yet this acceleration is accompanied by a growing chorus of references to “harm” that, upon closer inspection, reveal a striking lack of conceptual clarity. A generative model that proposes a high-performance composite may simultaneously accelerate sustainable energy storage and exacerbate resource depletion; an AI agent that optimizes synthesis pathways may reduce laboratory waste while inadvertently surfacing dual-use precursors suitable for chemical weapons; a large language model integrated with simulation tools may accelerate materials screening yet propagate subtle epistemic

distortions that mislead subsequent research programs [1-11]. In each case, the literature invokes “harm,” but rarely pauses to specify what kind of harm is at stake, through what causal pathway it arises, or why it merits distinct normative attention. The present paper contends that this ambiguity is not peripheral but central to the maturation of the field. When “harm” is used loosely across environmental, health, security, epistemic, social, and economic registers, researchers lose the capacity to differentiate mechanisms, compare severities, design targeted safeguards, or engage in coherent ethical deliberation [3, 12-19].

The problem manifests concretely across the literature. Butler and colleagues highlight the transformative promise of machine learning for molecular and materials science while nodding to “potential risks,” yet without disaggregating those risks into harm types [4]. Schmidt *et al.* [5] survey advances in solid-state materials modeling and mention unintended consequences of rapid discovery, again without taxonomic precision. Zunger’s influential account of inverse design celebrates target functionalities but leaves implicit the possibility that those targets may carry downstream harms once synthesized [6]. Montoya *et al.* [7] envision autonomous materials research yet acknowledge “future challenges” that remain unspecified in terms of harm. Parallel developments in AI ethics supply richer vocabularies—Floridi’s AI4People framework enumerates opportunities, risks, principles, and recommendations [2]; Hagendorff evaluates the limitations of guideline-based ethics [10]; Mittelstadt warns that principles alone cannot guarantee ethical AI [14]—yet these insights have not been systematically imported into materials-specific discourse [19-23]. Even recent contributions that explicitly address generative models in materials contexts, such as Persson and Montoya’s treatment of safety and security risks, or Spotte-Smith’s examination of ethical issues in large machine learning models, still employ “harm” in a manner that collapses distinct categories [22, 23].

This paper, therefore, intervenes at the conceptual boundary. It first surveys the inconsistent usages of “harm” in existing materials, AI and AI ethics literature, demonstrating that the term is pressed into service for at least five distinct phenomena. It then diagnoses three interlocking problems created by this conflation: category errors that obscure mitigation strategies, incommensurability that frustrates prioritization, and moral overload that renders the concept practically inert. Against this background, the paper advances a formal definition of scientific harm and a six-type typology calibrated to the materials AI domain. The typology is subsequently distinguished from neighboring terms, its internal relationships are mapped, common objections are addressed, and concrete implications for research practice are articulated. Throughout, the argument remains strictly conceptual and definitional, eschewing empirical claims, datasets, or performance metrics in favor of precise boundary-setting. The ultimate aim is to equip the materials AI community with a shared language capable of supporting harm-aware development without stifling innovation. By rendering “scientific harm” both intelligible and actionable, the framework seeks to ensure that the extraordinary capabilities of materials artificial intelligence are matched by equally sophisticated ethical and practical safeguards.

Harm in Existing Literature

A systematic examination of peer-reviewed publications from 2017 to 2024 reveals that “harm” functions as an umbrella term in materials AI discourse, absorbing at least five distinct usages without explicit differentiation. The first and perhaps most visible usage frames harm the environment [24-28]. Vinuesa *et al.* [12], for instance, explore artificial intelligence’s role in achieving the Sustainable Development Goals and note that AI-driven materials optimization can inadvertently intensify resource depletion or pollution if sustainability constraints are not foregrounded. Similarly, Bashir’s analysis of the climate and sustainability implications of generative AI highlights how training and inference costs, when coupled with materials discovery pipelines, may accelerate the extraction of rare earth elements or generate hazardous by-products at scales previously unimaginable [29]. Correa-Baena *et al.* describe high-throughput experimentation accelerated by machine learning, yet acknowledge that the resulting materials libraries may include compounds whose lifecycle environmental footprints remain unexamined [11].

A second usage centers on health harm. Urbina *et al.* [9] document dual-use concerns in artificial-intelligence-powered drug discovery, illustrating how generative models can propose molecular structures that, while therapeutically promising,

may also exhibit unforeseen toxicity profiles once synthesized and deployed [9]. McCarthy and Gupta extend this concern into materials science proper, arguing that ethical issues arise when AI systems recommend novel composites or alloys whose long-term human exposure risks are unknown [19]. The mechanism here is direct: the AI proposes a candidate, synthesis follows, and human or ecological contact produces adverse physiological outcomes.

Security harm constitutes a third prominent register. Grinbaum and Adomaitis [8] foreground dual-use concerns of generative AI and large language models, warning that materials design tools could be repurposed to engineer high-energy-density compounds suitable for military applications. Urbina *et al.* [9] provide a concrete parallel in the chemical domain, noting that generative models capable of proposing novel molecules for legitimate research may simultaneously lower barriers to chemical-weapon precursors. Persson and Montoya focus explicitly on safety and security risks from generative materials models, emphasizing how autonomous systems might surface structures with unintended weaponization potential [22]. Cave's framework for responsible innovation in military AI contexts further underscores the ease with which materials AI outputs can migrate into security-sensitive domains [20].

Epistemic harm appears as a fourth, more subtle usage. The self-referential contribution by Cocito *et al.* [3] already isolates "scientific harm" as damage inflicted upon knowledge systems themselves. Hagendorff's evaluation of AI ethics guidelines reveals how over-reliance on opaque models can erode scientific trust when predictions prove systematically misleading [10]. Mittelstadt similarly cautions that principles alone cannot prevent epistemic distortions when AI systems shape research agendas [14]. Spotte-Smith extends this line of reasoning to large machine learning models in materials science, observing that hallucinated property predictions or overfitted training sets can propagate false positives that waste downstream experimental effort and distort collective understanding [23].

Finally, social harm surfaces in sociotechnical analyses. Shelby and colleagues develop a taxonomy of sociotechnical harms of algorithmic systems that readily maps onto materials AI, noting how unequal access to high-performance computing resources can concentrate the benefits of accelerated discovery among well-resourced institutions while marginalizing others [16]. Coeckelbergh and Leslie each emphasize that AI ethics must attend to distributive justice; when material innovations disproportionately benefit affluent markets, existing inequalities are amplified [17, 18]. McCarthy and Gupta again bridge the gap to materials science by questioning whether AI-driven discovery exacerbates global disparities in access to advanced materials [19].

These five usages—environmental, health, security, epistemic, and social—are not exhaustive but sufficiently representative to illustrate the literature's current state. Notably absent is any sustained attempt to relate the categories to one another or to derive mitigation strategies that respect their distinct causal pathways. Bostrom's foundational treatment of superintelligence risks supplies an early precedent for considering large-scale harms [1]. At the same time, Floridi's AI4People framework offers a multi-dimensional ethical scaffold [2], yet neither has been translated into materials-specific typologies. Jobin *et al.* map the global landscape of AI ethics guidelines [13] and Morley *et al.* [15] review tools for translating principles into practice, but again without differentiation of harm types. Dignum's account of AI ethics in responsible innovation [21] and Undheim's governance analysis of AI-enabled synthetic biology [28] further enrich the conceptual space without resolving the terminological ambiguity that persists in materials AI proper. The result is a literature rich in warnings yet conceptually under-specified, precisely the lacuna the present typology seeks to address.

The Problem with Current Usage

The conflation of disparate harm types under a single umbrella term generates three interlocking conceptual and practical problems that undermine both scientific integrity and ethical responsibility in materials AI.

The first problem is a category error. When a paper asserts that an AI model "may cause harm" without specifying the type, readers cannot determine whether the concern involves ecosystem disruption from solvent-intensive synthesis pathways recommended by the model, physiological damage from exposure to a newly proposed nanomaterial, or erosion of trust in published results because the model systematically overestimates the stability of metastable phases

[22, 23]. Environmental harms operate through biophysical mechanisms and are mitigated by lifecycle analysis and green chemistry constraints. Health harms require toxicological profiling and regulatory oversight. Security harms demand export-control regimes and access restrictions. Epistemic harms call for reproducibility standards and uncertainty quantification. Because these mechanisms differ, a mitigation strategy effective for one type—say, carbon-footprint reporting for environmental harm—offers no protection against another—say, dual-use screening for security harm [8, 9]. Treating them as interchangeable, therefore, produces mismatched safeguards and false reassurance.

The second problem is incommensurability. Without typed distinctions, it becomes impossible to compare or trade off harms in any principled manner. Is the environmental cost of training a generative materials model commensurate with the epistemic benefit of faster discovery? Is a hypothetical security risk from a dual-use compound more or less tolerable than documented social harms arising from unequal access to AI tools [12, 16, 29]? The literature currently provides no metric or even vocabulary for such comparisons. Floridi and colleagues rightly insist that ethical frameworks must weigh opportunities against risks [2], yet without disaggregation of harm types, the weighing exercise collapses into a rhetorical gesture. Hagendorff's critique of guideline proliferation makes the same point: principles remain aspirational unless harms are rendered comparable [10].

The third and perhaps most insidious problem is moral overload. When "harm" is stretched to cover every conceivable negative outcome, the term loses prescriptive force. Researchers may acknowledge "potential harms" in the discussion section as a ritualistic nod to ethics without altering experimental design, data curation, or dissemination practices [14, 19]. Mittelstadt warns that principles alone cannot guarantee ethical AI precisely because vague language permits superficial compliance [14]. In materials AI, this manifests as papers that list "environmental and societal concerns" yet proceed with resource-intensive hyperparameter searches or publish candidate structures without accompanying harm pathway analysis [11, 24]. The overload also affects reviewers, who lack clear criteria for evaluating harm claims, and funding bodies, who cannot demand proportionate safeguards. The cumulative effect is ethical fatigue: the language of harm, once potent, becomes background noise.

These three problems—category error, incommensurability, and moral overload—are not abstract philosophical worries. They directly impair the capacity of the materials AI community to fulfill its responsibility to anticipate, characterize, and mitigate the negative consequences of its own tools [3, 21]. A boundary definition and typology are therefore not optional refinements but necessary preconditions for coherent practice.

Proposed Definition of Scientific Harm

To resolve the ambiguities diagnosed above, this paper advances the following formal definition:

Scientific harm in the context of misused materials AI is any negative consequence—whether direct, indirect, potential, or actual—arising from the development, deployment, or misuse of artificial intelligence systems whose primary purpose or output concerns the discovery, design, characterization, or optimization of materials, where that consequence foreseeably and avoidably affects humans, non-human environments, scientific knowledge systems, or social structures in ways that violate legitimate normative expectations within the scientific enterprise.

The definition is deliberately scoped to materials AI to preserve disciplinary precision while remaining broad enough to encompass the full spectrum of impacts. It excludes harms unrelated to materials (for example, bias in general-purpose large language models) and focuses on consequences that are causally traceable to the AI intervention rather than to downstream human decisions alone.

Four distinctions internal to the definition merit elaboration. Direct harm occurs along an immediate causal pathway: an AI-recommended synthesis route produces a toxic intermediate that exposes laboratory personnel before any human oversight can intervene [9]. Indirect harm is mediated: an AI-optimized composite enters global supply chains, whose extraction practices then degrade distant ecosystems [29]. Potential harm describes a risk that has not yet materialized

but is rendered more probable by the AI system, such as the increased accessibility of dual-use precursors [8, 22]. Actual harm denotes a realized negative outcome, whether already observed or retroactively traceable.

The definition further requires that the consequence be both foreseeable (given reasonable scientific diligence) and avoidable (through feasible design choices). This normative clause prevents the trivialization of harm as an inevitable byproduct of progress. Bostrom's analysis of superintelligence paths underscores the importance of anticipatory governance precisely because some harms become foreseeable long before they become actual [1]. Floridi's framework similarly insists that ethical AI must weigh avoidable harms against benefits [2]. By embedding foreseeability and avoidability, the definition supplies a test that materials scientists can apply when evaluating their own models and outputs.

Importantly, the definition is not outcome-monistic; it accommodates plural normative expectations—scientific reliability, environmental integrity, public health, global security, and social justice—without privileging any single axis. This pluralism is essential because materials AI operates at the intersection of epistemic, material, and societal domains [3, 16, 19]. The definition, therefore, functions as a boundary concept that delimits the phenomenon under study while inviting the subsequent typology to articulate its internal structure.

Figure 1 translates the proposed definition into a hierarchical decision structure that shows how misused materials AI becomes legible as scientific harm only when normatively screened and then differentiated into six analytically distinct harm types.

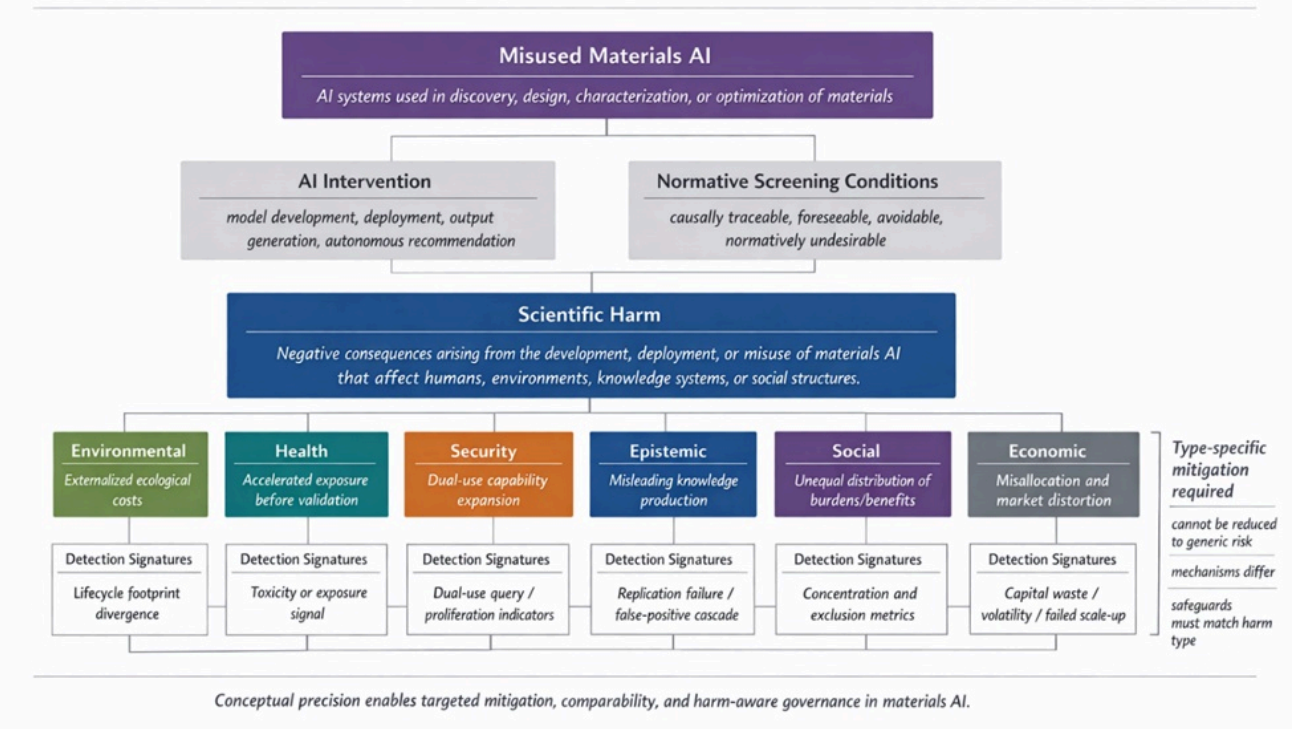

Figure 1. Hierarchical decision structure of scientific harm in misused materials AI

A Typology of Scientific Harm

The proposed definition is operationalized through a six-type typology calibrated to the materials AI domain. Each type is defined, its causal mechanism is explicated, materials-specific examples are furnished, and detectable signatures are

identified to facilitate empirical monitoring.

Environmental harm is damage to ecosystems, resource depletion, or pollution attributable to AI-recommended materials or the processes used to discover and deploy them. The mechanism involves AI systems optimizing for performance metrics while externalizing environmental costs—training compute emissions, solvent consumption, or rare-element extraction [12, 29]. Example: a generative model proposes a high-capacity battery cathode whose synthesis requires cobalt mined under ecologically destructive conditions; the model accelerates discovery but does not constrain feedstock sustainability. Detection signature: lifecycle assessment discrepancies between predicted and actual environmental footprints, or spikes in Scope 3 emissions correlated with AI-generated candidate libraries [11].

Health harm consists of adverse physiological effects on humans or non-human biota resulting from exposure to materials whose discovery or optimization was materially assisted by AI. The mechanism is the acceleration of synthesis and deployment before comprehensive toxicological data accumulate [9, 19]. Example: an inverse-design tool identifies a novel polymer with superior mechanical properties that later proves to release microplastics or endocrine-disrupting leachates upon degradation. Detection signature: post-market epidemiological signals or in-vitro assays that diverge from the AI's property predictions.

Security harm arises when AI systems facilitate the creation or proliferation of materials that enable weapons, surveillance, or other malicious applications. The mechanism is the removal of tacit knowledge barriers; generative models democratize access to high-performance energetic materials or stealth coatings [8, 9, 22]. Example: a diffusion model trained on open crystallographic data proposes metastable high-explosive candidates that evade existing detection protocols. Detection signature: sudden increases in queries for dual-use precursors or publication patterns that match known proliferation indicators [20].

Epistemic harm is damage inflicted upon scientific knowledge systems, including the propagation of misleading results, the waste of collective research effort, and the erosion of warranted trust. The mechanism is the introduction of systematic biases, hallucinations, or overconfidence in AI outputs that shape downstream experimentation [3, 10, 23]. Example: a large language model integrated with density-functional theory pipelines consistently underestimates formation energies, leading an entire subfield to pursue metastable phases that prove unrealizable. Detection signature: replication failures, citation cascades built on retracted or corrected AI-derived claims, or meta-analyses showing divergence between predicted and measured properties.

Social harm encompasses the inequitable distribution of benefits and burdens, the exacerbation of existing inequalities, or the displacement of communities linked to AI-accelerated materials innovation. The mechanism is the concentration of computational resources and proprietary datasets among elite institutions, coupled with the globalized supply chains that materials AI helps optimize [16, 17, 19]. Example: AI-designed photovoltaics dramatically lower costs in high-income markets while the requisite critical minerals are extracted from regions experiencing land dispossession. Detection signature: bibliometric imbalances, patent concentration metrics, or stakeholder reports documenting livelihood disruption.

Economic harm involves resource waste, market disruption, or unfair competitive distortions produced by AI systems in materials contexts. The mechanism is the misallocation of capital toward AI-generated candidates that fail at scale or the premature market entry of materials that render prior investments obsolete [24, 26]. Example: venture-funded start-ups pour resources into scaling AI-proposed alloys that later prove economically unviable because the model ignored supply-chain fragility. Detection signature: high failure rates of AI-derived patents, sudden market-share volatility in materials sectors, or investor reports citing "over-optimistic AI forecasts."

The six types are not mutually exclusive; a single AI system may generate harms across multiple categories simultaneously. The typology's value lies in furnishing a shared vocabulary that replaces vague assertions of "risk" with typed, mechanistically grounded claims. By requiring authors to locate their concerns within this framework, the community gains the precision necessary for cumulative progress in harm-aware materials AI.

Table 1 consolidates the six harm types into an analytical matrix that clarifies what each type affects, how it emerges, and which mitigation logic is most appropriate.

Table 1. Cross-type analytical matrix for scientific harm in misused materials AI

Harm type	Primary object affected	Typical AI-mediated causal mechanism	Temporal profile	Main level of manifestation	What makes this type analytically distinct	Most appropriate first-line mitigation logic
Environmental	Ecosystems, resource stocks, and emissions pathways	Optimization for technical performance while ecological costs remain external to model objectives	Often delayed and accumulative	Supply chain, lifecycle, planetary systems	The central issue is biophysical degradation rather than human toxicity, knowledge distortion, or distributive inequity.	Embed lifecycle constraints, resource provenance checks, and sustainability-aware objective functions.
Health	Human and non-human bodies	Rapid proposal and deployment of poorly characterized materials before sufficient toxicological validation	Can be immediate or delayed	Laboratory, product exposure, and environmental uptake	The defining concern is physiological injury or exposure harm rather than ecological footprint or knowledge failure.	Require toxicology screening, exposure assessment, phased validation, and post-deployment monitoring.
Security	Public safety, strategic stability, and controlled knowledge	Lowering tacit knowledge barriers to dual-use materials design or weapon-relevant discovery	Often latent until the misuse pathway is activated	National security, defense, and proliferation networks	Distinct because the core issue is malicious enablement, not ordinary performance failure or inequality alone.	Use access controls, dual-use review, red-teaming, dissemination limits, and escalation protocols.
Epistemic	Scientific knowledge systems, trust, and	Hallucinations, bias, overconfidence, data leakage, or non-reproducible	Often cumulative and self-reinforcing	Publication system, benchmarking, and field-wide inference	This type is unique because it damages the reliability and	Strengthen uncertainty reporting, reproducibility standards,

	research agendas	predictions shape downstream inquiry			credibility of knowledge itself.	benchmark discipline, and independent validation.
Social	Communities, institutions, distributive structures	Unequal access to computing, datasets, patents, and benefits from AI-enabled materials innovation	Medium- to long-term	Institutions, regions, labor systems, and communities	The issue is patterned inequality in burdens and benefits rather than direct toxicity or market mispricing.	Use distributional assessment, inclusive governance, stakeholder review, and benefit-sharing mechanisms.
Economic	Capital allocation, firms, markets, innovation portfolios	Overinvestment in weak AI-derived candidates, distorted expectations, premature scaling, and ignored supply fragility	Often appears after translation or commercialization	Firms, investors, and industrial ecosystems	It is analytically distinct because the primary damage lies in wasted resources, instability, or unfair competitive distortion.	Require techno-economic validation, staged investment gates, supply-chain stress testing, and market realism checks.

Distinctions From Nearby Terms

To sharpen the proposed definition and typology, scientific harm must be explicitly distinguished from five neighboring concepts frequently conflated with it in materials AI discourse. A conceptual comparison table—structured along four distinguishing dimensions (causal immediacy, normative valence, foreseeability, and mitigability)—clarifies these boundaries in full sentences. First, harm differs from risk [2, 14] because risk denotes a probability-weighted possibility of a negative outcome, whereas harm is the realized or actualized negative consequence itself; for example, an AI model that increases the probability of producing a toxic material poses a risk, but the actual synthesis and exposure constitute harm [9, 19]. Second, harm is distinct from danger [1, 22] in that danger refers to an inherent potential for harm latent in a material or system, while harm requires the actualization of that potential through AI-mediated pathways; thus, a generative materials model may highlight a dangerous dual-use compound, yet only its deployment creates security harm [8]. Third, harm departs from misuse [3, 8] because misuse implies intentional harmful application by a human actor, whereas harm can arise unintentionally from otherwise legitimate AI outputs; an inverse-design tool employed for benign battery research may still generate epistemic harm if its predictions mislead the field without any malicious intent. Fourth, harm contrasts with unintended consequence [5, 7] in that unintended consequences may be positive or neutral, but harm is normatively negative and causally traceable to the AI intervention; accelerating discovery via AI might yield an unintended efficiency gain (positive) or an unintended pollution pathway (harm). Fifth, harm is not identical to side effects [4, 11] because side effects are collateral outcomes that need not violate normative expectations, whereas harm specifically breaches legitimate scientific or societal standards; a side effect of faster screening might be minor computational overhead, but the same process producing environmental harm crosses the threshold. Along the causal-immediacy dimension, harm is more immediate than risk or danger; along normative valence, harm carries inherent negative weight absent in unintended consequences; foreseeability is higher for harm under the proposed definition than

for vague side effects; and mitigability requires type-specific strategies unavailable when terms remain undifferentiated [10, 16]. These distinctions prevent the category errors diagnosed earlier and equip practitioners to speak with precision.

Table 2 clarifies the manuscript's conceptual boundary by distinguishing scientific harm from adjacent terms that are often used interchangeably in materials AI discourse.

Table 2. Boundary conditions distinguishing scientific harm from adjacent concepts in materials AI

Concept	Core definition in this manuscript	Normative valence	Relationship to AI-mediated materials research	Causal status	Why is it not identical to scientific harm	Materials AI example
Scientific harm	A negative consequence arising from the development, deployment, or misuse of materials AI that is causally traceable, foreseeable, avoidable, and normatively undesirable within scientific practice.	Intrinsically negative	Directly anchors the manuscript's analytical object.	Can be potential or actual, but must be specified as a consequence pathway.	This is the focal category against which neighboring concepts are differentiated.	An AI-derived materials pipeline produces misleading stability predictions that redirect large amounts of experimental effort toward unrealizable compounds.
Risk	A probability-weighted possibility that a negative outcome may occur.	Prospectively negative but not yet realized	Often appears in materials AI discussions as a forward-looking concern about future consequences.	Ex ante condition of uncertainty.	Risk concerns the chance of harm, whereas scientific harm denotes the negative consequence itself once analytically specified as a traceable outcome.	A generative model increases the likelihood that toxic candidate materials will be proposed.
Danger	An inherent potential for harm latent in a material, process, or model capability.	Potentially negative	Refers to the hazardous capacity embedded in an output or system, whether or not downstream	Latent property or condition.	Danger can exist without any actual harmful pathway being activated, whereas scientific harm requires AI-	A model surfaces a high-energy compound with weaponization potential, but no harmful use yet occurs.

			consequences occur.		mediated causal realization or a clearly articulated harmful consequence.	
Misuse	Intentional or negligent use of a system for purposes outside legitimate scientific aims.	Negative because of the actor conduct	Highlights human agency in applying AI outputs in problematic ways.	Behavioral and intentional pathway.	Misuse is one route through which scientific harm may arise, but harm can also occur unintentionally through ordinary research use.	A benign design model is deliberately used to search for militarily relevant energetic materials.
Unintended consequence	Any outcome not originally sought by the actor or designer.	Normatively open: can be positive, neutral, or negative	Captures the surprise effects of AI deployment in research environments.	Consequential but not normatively fixed.	Not every unintended consequence is harmful; scientific harm is a subset marked by negative normative significance and traceable damage.	AI screening unexpectedly reduces lab waste, or alternatively accelerates pollution-intensive synthesis choices.
Side effect	A collateral outcome accompanying the primary intended effect of a system or intervention.	May be trivial, neutral, or negative	Common in technical descriptions of trade-offs or spillovers from optimization.	Secondary outcome	A side effect becomes scientific harm only when it crosses a threshold of normatively undesirable consequence affecting people, environments, knowledge, or social structures.	Increased compute demand during model training may be a side effect; it becomes environmental harm when it materially worsens emissions or extraction burdens.

Relationship between Harm Types

The six harm types do not exist in isolation; they interlink through causal, compounding, and feedback relationships that the typology makes visible. Four primary relationships merit articulation. First, environmental harm frequently precipitates health harm: AI-optimized extraction processes deplete ecosystems and generate pollutants that enter human exposure pathways, as when cobalt-mining by-products from battery materials discovery contaminate water supplies [12, 29]. Second, security harm can cascade into social harm: dual-use materials enabling surveillance technologies exacerbate inequalities when access is restricted to powerful actors, displacing communities or widening global divides [8, 20]. Third, epistemic harm readily leads to economic harm: misleading AI predictions waste capital on scaling unviable candidates, producing market distortions and investor losses [3, 23, 26]. Fourth, harms compound multiplicatively; a single generative model may simultaneously produce environmental harm (resource intensity), epistemic harm (hallucinated properties), and social harm (concentrated benefits), creating synergistic effects greater than any isolated type [9, 22]. These relationships can be visualized in a conceptual figure: six nodes arranged in a hexagon (environmental, health, security, epistemic, social, economic), with directed arrows showing unidirectional and bidirectional flows—e.g., a thick arrow from environmental to health labeled “pollution-exposure pathway,” a dashed arrow from epistemic to economic labeled “misallocation feedback,” and a central hub indicating compounding at the AI-system level. The figure underscores that mitigation must address relational dynamics rather than isolated types, echoing Floridi’s multi-dimensional framework while grounding it in materials-specific mechanisms [2, 19].

Objections and Replies

Three common objections arise against the typology; each receives a targeted reply grounded in the framework’s conceptual resources. Objection 1 holds that harm is inherently subjective, with stakeholders differing on what counts as negative. The reply is that the typology accommodates perspectival pluralism by anchoring types in traceable causal pathways and shared normative expectations within scientific practice; subjectivity does not preclude systematic analysis, as Bostrom’s risk taxonomy demonstrates [1, 3]. Objection 2 claims the framework is speculative because most harms remain hypothetical in early-stage materials AI. The reply counters that anticipation is the point: early identification enables prevention, precisely as Hagendorff and Mittelstadt argue when critiquing reactive ethics [10, 14]. Objection 3 asserts that materials AI is too nascent for such analysis, risking over-regulation that stifles innovation. The reply is that early-stage intervention is most effective; defining harm now shapes development before entrenched practices make mitigation costlier, aligning with Dignum’s responsible-innovation call and the seed literature’s own forward-looking stance [7, 21]. These replies preserve the typology’s utility without dismissing legitimate concerns.

Implications for Materials AI Practice

The typology and definition carry immediate implications for three stakeholder groups. For authors, three practices become mandatory: (a) specify the harm type when discussing risks, linking each claim to a detectable signature and causal mechanism; (b) provide explicit evidence of harm pathways rather than generic disclaimers; and (c) outline type-specific mitigations, such as green-chemistry constraints for environmental harm or dual-use screening for security harm [22, 23]. For reviewers, two obligations follow: (a) require harm-type specification in manuscripts and (b) verify consistency between claimed harms and proposed safeguards. For the broader community, three collective actions are indicated: (a) adopt harm-reporting standards in journal guidelines, (b) develop shared harm-assessment templates calibrated to the six types, and (c) fund studies of cross-type mitigation strategies. These changes transform vague ethical nods into precise, actionable governance, fulfilling the promise of Floridi’s and Jobin’s ethics roadmaps within materials contexts [2, 13].

Conclusion

This boundary paper has identified the ambiguous usage of “harm” across materials AI literature, surveyed five distinct usages, diagnosed the problems of conflation, proposed a formal definition of scientific harm, and advanced a six-type

typology—environmental, health, security, epistemic, social, and economic—complete with mechanisms, examples, and signatures. It has distinguished harm from risk, danger, misuse, unintended consequence, and side effect; mapped inter-type relationships; answered key objections; and outlined practice implications. The typology equips the field to replace overloaded rhetoric with precise, relational analysis. Future work can refine detection signatures and test the framework empirically. Still, the conceptual boundary is now set: precise language about typed harms is essential if materials artificial intelligence is to realize its transformative potential without inflicting avoidable damage on science, society, or the planet.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 29 Jul 2023 Revised: 03 Oct 2023 Accepted: 15 Nov 2023
Published online: 18 January 2024

References

Kim H, Yi X, Yao J, Lian J, Huang M, Duan S, et al. The road to artificial superintelligence: A comprehensive survey of superalignment. arXiv [Preprint]. 2024:arXiv:2412.16468.

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.

Cocito C, Marquenie T, De Hert P. Risk, harm and damage as preset rational categories in ai literature: Do we see or think the problem? Eur J Law Technol. 2024;15(3).

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.

Grinbaum A, Adomaitis L. Dual use concerns of generative AI and large language models. J Responsible Innov. 2024;11(1):2304381.

Urbina F, Lentzos F, Invernizzi C, Ekins S. Dual use of artificial-intelligence-powered drug discovery. Nat Mach Intell. 2022;4(3):189-91.

Hagendorff T. The ethics of AI ethics: An evaluation of guidelines. Minds Mach. 2020;30(1):99-120. https://doi.org/10.1007/s11023-020-09517-8.

Correa-Baena JP, Hippalgaonkar K, Van Duren J, Jaffer S, Chandrasekhar VR, Stevanovic V, et al. Accelerating materials development via automation, machine learning, and high-performance computing. Joule. 2018;2(8):1410-20.

Vinuesa R, Azizpour H, Leite I, Balaam M, Dignum V, Domisch S, et al. The role of artificial intelligence in achieving the sustainable development goals. Nat Commun. 2020;11(1):233.

Jobin A, Ienca M, Vayena E. The global landscape of AI ethics guidelines. Nat Mach Intell. 2019;1(9):389-99.

Mittelstadt B. Principles alone cannot guarantee ethical AI. Nat Mach Intell. 2019;1(11):501-7.

Morley J, Floridi L, Kinsey L, Elhalal A. From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. In: Ethics, governance, and policies in artificial intelligence. Cham: Springer International Publishing; 2021. p. 153-83.

Shelby R, Rismani S, Henne K, Moon A, Rostamzadeh N, Nicholas P, et al. Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In: Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. New York: Association for Computing Machinery; 2023. p. 723-41.

Bostrom N, Yudkowsky E. The ethics of artificial intelligence. In: Artificial intelligence safety and security. Boca Raton: Chapman and Hall/CRC; 2018. p. 57-69.

Leslie D. Understanding artificial intelligence ethics and safety. arXiv [Preprint]. 2019:arXiv:1906.05684.

Karimian G, Petelos E, Evers SM. The ethical issues of the application of artificial intelligence in healthcare: A systematic scoping review. AI Ethics. 2022;2(4):539-51.

Stanley-Lockman Z. Responsible and ethical military AI. Washington, DC: Centre for Security and Emerging Technology; 2021.

Herrmann H. What's next for responsible artificial intelligence: A way forward through responsible innovation. Heliyon. 2023;9(3):e14379.

Janjeva A, Harris A, Mercer S, Kasprzyk AM, Gausen A. The rapid rise of generative AI: Assessing risks to safety and security. London: Centre for Emerging Technology and Security; 2023. Available from: https://cetas.turing.ac.uk/sites/default/files/2023-12/cetas_research_report_-_the_rapid_rise_of_generative_ai_-_2023.pdf

Venkatasubbu S, Krishnamoorthy G. Ethical considerations in AI addressing bias and fairness in machine learning models. J Knowl Learn Sci Technol. 2022;1(1):130-8.

Juan Y, Dai Y, Yang Y, Zhang J. Accelerating materials discovery using machine learning. J Mater Sci Technol. 2021;79:178-90.

Wang Y, Wang K, Zhang C. Applications of artificial intelligence/machine learning to high-performance composites. Compos B Eng. 2024;285:111740.

Menon D, Ranganathan R. A generative approach to materials discovery, design, and optimization. ACS Omega. 2022;7(30):25958-73.

Aykol M. Materials discovery through artificial intelligence. Bull Am Phys Soc. 2020;65(1):M39.00001.

Undheim TA. The whack-a-mole governance challenge for AI-enabled synthetic biology: Literature review and emerging frameworks. Front Bioeng Biotechnol. 2024;12:1359768.

Bashir N, Donti P, Cuff J, Sroka S, Ilic M, Sze V, et al. The climate and sustainability implications of generative AI. Cambridge: MIT Press; 2024. Available from: https://mit-genai.pubpub.org/pub/8ulgrckc/release/2