ORIGINAL RESEARCH Open access

Conceptual Foundations for Value-Sensitive Design of Materials AI Systems

Claire Dupont[1*], Julien Martin[1]

Abstract

Materials AI systems, which apply machine learning techniques to accelerate the discovery, design, and optimization of new materials, are not value-neutral despite frequent claims of technical objectivity. Every design choice—ranging from the selection of training datasets and the formulation of objective functions to the prioritization of target properties and the definition of success metrics—necessarily embeds specific values, even when researchers present their work as purely data-driven or performance-oriented. Yet these embedded values remain largely unexamined in the current literature on materials informatics and autonomous materials research. The values at stake span multiple dimensions: epistemic values such as accuracy, reproducibility, and interpretability that underpin scientific validity; ethical values including safety, fairness, and accountability to prevent harm; social values like equity and benefit-sharing that address who gains from new materials; economic values focused on efficiency, cost-effectiveness, and scalability; and environmental values centered on sustainability, non-toxicity, and circularity. Value-sensitive design (VSD), originally developed within human-computer interaction as a principled approach to technology development, provides a systematic framework for making these values visible, negotiable, and actionable rather than leaving them implicit or unacknowledged. Building directly on established VSD foundations, this paper proposes a conceptual framework tailored specifically to materials AI contexts. The framework includes five integrated components—value identification, stakeholder mapping, value operationalization, design translation, and value evaluation—alongside a typology of five value categories and explicit guidance for navigating common value tensions. By adapting VSD principles to the unique challenges of materials discovery, the framework offers a pathway to responsible innovation that aligns technical capabilities with broader human and environmental priorities. Ultimately, adopting value-sensitive practices in materials AI will help ensure that AI-augmented materials research contributes not only to faster discovery but also to more equitable, sustainable, and ethically sound outcomes for society and the planet.

Keywords Materials AI, Responsible innovation, Value-sensitive design, Ethics in artificial intelligence, Sustainability in materials science, Stakeholder values

*Correspondence:
Claire Dupont
claire.dupont@gmail.com
[1] Department of AI-Based Materials Engineering, University of Bordeaux, Bordeaux, France

Introduction

Materials AI systems have emerged as a transformative force in materials science, promising to accelerate discovery, reduce experimental costs, and unlock novel functionalities through data-driven approaches. Yet beneath the surface of these technical achievements lies a fundamental oversight: materials AI systems are not value-neutral. Choices about what to optimize, which data to collect, which properties to prioritize, and which metrics define success are never purely technical; they inevitably embed specific values that shape both the process and the outcomes of research. For instance, an algorithm trained predominantly on high-performance structural alloys may implicitly prioritize economic efficiency over environmental sustainability, while a generative model focused on rapid screening may sideline considerations of long-term safety or equitable access to resulting innovations. These value-laden decisions occur at every stage—from data curation and model architecture to evaluation protocols—yet they are rarely discussed explicitly in the literature [1-4].

This paper argues that the current neglect of values in materials AI is not merely an oversight but a structural feature of the field's development. Technical excellence in prediction accuracy, computational speed, and novelty continues to dominate research agendas, while normative assumptions remain unexamined. The consequences are significant: unacknowledged values can lead to materials that exacerbate environmental degradation, reinforce social inequities, or introduce unforeseen safety risks, all while being presented as objective scientific progress [5-7]. To address this gap, the present work introduces value-sensitive design (VSD) as a conceptual framework capable of making values visible and actionable within materials AI. Originally developed for information systems, VSD provides a structured methodology for identifying stakeholders, eliciting their values, and translating those values into concrete design decisions [4-9].

By adapting VSD to the materials science context, this article demonstrates that values are already embedded in existing materials AI pipelines, even when designers claim neutrality. It further articulates why such values have been neglected, presents an adapted VSD framework with five core components, offers a typology of five value categories particularly relevant to materials AI, and analyzes key value tensions that arise in practice. The overall aim is to shift materials AI from a narrowly technical enterprise toward one that explicitly serves human and environmental well-being. In doing so, the framework aligns with broader calls for responsible innovation in scientific AI, ensuring that the powerful tools now reshaping materials discovery are guided by deliberate, transparent, and democratically informed value commitments rather than defaulting to unexamined assumptions.

What Are Values in Materials AI?

Values in materials AI can be understood as the principles, priorities, or qualities that shape how systems are designed, deployed, and evaluated, extending beyond explicit objectives to include the implicit assumptions that govern what is optimized, what is measured, and what is ultimately produced. This framing situates values at the intersection of epistemic, ethical, social, economic, and environmental dimensions, reflecting the fact that materials AI does not operate within a purely technical vacuum but within a broader landscape of scientific practice and societal consequence. Under this view, values are not external constraints imposed after the fact; they are embedded within the very structure of computational workflows [10-12].

This embeddedness becomes visible when examining seemingly technical decisions that, in practice, encode normative commitments. The selection of molecular descriptors, for instance, is not a neutral preprocessing step but a decision about which aspects of material reality are deemed relevant for representation. Similarly, the construction of an objective function that prioritizes predicted bandgap while neglecting synthesis feasibility implicitly elevates theoretical performance over manufacturability, thereby privileging one form of success over another. Such choices illustrate that value judgments are not exceptional interventions but routine features of model construction [9-14].

The manifestation of values extends across the full pipeline of materials AI, revealing a layered structure in which normative considerations are distributed rather than localized. Decisions at the data level determine which materials histories are preserved and which are excluded, shaping the evidentiary base upon which models are trained. At the modeling level, architectural preferences—such as the adoption of black-box neural networks over interpretable alternatives—reflect a prioritization of predictive capacity relative to transparency. Evaluation practices further reinforce these orientations, as metrics that reward rapid discovery without accounting for long-term stability or safety implicitly privilege short-term gains over sustained reliability. These layered manifestations are not merely theoretical; they materialize in systems that accelerate the identification of high-efficiency catalysts while leaving questions of environmental persistence or toxicity insufficiently addressed [13-17].

Yet the presence of values does not imply uniformity. On the contrary, materials AI is characterized by a plurality of competing priorities that reflect the diverse interests of its stakeholders. Academic researchers may emphasize explanatory depth, industry actors may prioritize scalability and performance, regulatory institutions may focus on safety and compliance, and environmental perspectives may foreground sustainability. These orientations do not readily align, and the design of any given system necessarily involves trade-offs among them. Recognizing this plurality challenges the notion that technical decisions can remain insulated from normative considerations and instead points toward a more reflexive mode of practice in which such commitments are explicitly articulated and justified.

Why Values Are Currently Neglected

Despite their pervasive presence, values remain largely unexamined within contemporary materials AI research. This condition arises not from their absence but from a set of reinforcing dynamics within the field. One such dynamic is the prevailing emphasis on technical performance, where advances are primarily evaluated through improvements in accuracy, efficiency, or the novelty of discovered materials. Within this evaluative framework, normative analysis appears secondary, if not extraneous, leading researchers to concentrate on problems that are readily quantifiable while leaving underlying value structures implicit.

This orientation is further sustained by a persistent belief in the neutrality of data-driven methods. Algorithms are often framed as instruments that reveal patterns inherent in nature, a characterization that obscures the multitude of decisions embedded within their construction. The language of objectivity commonly associated with machine learning reinforces this perception, rendering invisible the value-laden processes of data curation, feature selection, and objective formulation. When systems are treated as neutral, the assumptions they carry are neither scrutinized nor contested.

The difficulty of engaging with values is compounded by the absence of formal training in relevant domains. Educational pathways in materials science and artificial intelligence typically prioritize technical competencies—mathematical modeling, computational implementation, and domain-specific knowledge—while offering little exposure to ethics, philosophy of science, or frameworks for value analysis. As a result, practitioners may lack both the conceptual vocabulary and methodological tools required to identify and interrogate the normative dimensions of their work [18-20].

These challenges are reinforced at the institutional level through incentive structures that rarely reward such reflection. Academic publishing, peer review processes, and funding mechanisms continue to privilege methodological rigor and performance metrics, often without requiring explicit consideration of the values embedded in research design. Even when broader impacts are invoked, they are frequently framed in terms of technological readiness or economic benefit rather than as sites of normative deliberation. In this context, the omission of value analysis is not an oversight but a rational response to prevailing expectations.

The interaction among these dynamics produces a self-reinforcing cycle in which technical focus marginalizes normative inquiry, assumptions of neutrality render it unnecessary, limited training constrains its articulation, and institutional incentives discourage its pursuit. Interrupting this cycle requires more than incremental adjustment; it demands both conceptual frameworks capable of making values visible and structural changes that legitimize their examination as an integral component of scientific practice.

Value-Sensitive Design: Origins and Principles

Value-Sensitive Design originated in the field of human-computer interaction as a principled methodology for incorporating human values into technology development from the earliest stages. It was developed to counteract the tendency of designers to focus exclusively on technical functionality while neglecting the broader ethical and social implications of their creations. The approach rests on three interrelated phases that together ensure values are not afterthoughts but core design considerations.

The conceptual phase involves identifying key stakeholders and the values they hold. Stakeholders include direct users, indirect beneficiaries, and those potentially harmed by the technology; values are elicited through philosophical analysis, empirical investigation, and normative reflection. The empirical phase examines how well current or proposed designs support or violate those identified values, often through interviews, observations, or prototypes. Finally, the technical phase translates value commitments into concrete system features, algorithms, and evaluation criteria [21-24].

When adapted to materials AI, these three phases take on distinctive characteristics. In the conceptual phase, stakeholders extend beyond human users to include future generations affected by material legacies and even the environment itself as a non-human stakeholder whose “interests” (sustainability, biodiversity) must be represented. The empirical phase must account for the long time horizons of material deployment, where safety or environmental impacts may only become evident years or decades later. The technical phase requires translating abstract values such as sustainability into specific, computable constraints within generative models, surrogate functions, or multi-objective optimization routines.

This adaptation preserves VSD’s core commitment that values and technology co-evolve: technical choices shape value realization, while explicit value commitments guide technical refinement. Applied to materials AI, VSD thus offers a way to move from implicit value embedding to deliberate value alignment [25-27].

A Framework for Value-Sensitive Materials AI

This paper proposes a conceptual framework for value-sensitive design in materials AI that seeks to render normative commitments explicit and technically actionable. Rather than treating values as external constraints or post hoc considerations, the framework positions them as integral to the architecture of materials AI systems, shaping decisions from initial problem formulation through to evaluation and deployment. The aim is not simply to acknowledge the presence of values, but to operationalize them in a manner that allows systematic incorporation into scientific and computational practice.

The process begins with the deliberate identification of values that are relevant to a given material’s AI task, recognizing that these are often only partially visible within conventional technical formulations. Drawing on domain-specific literature as well as broader conceptual typologies, this stage surfaces both explicit priorities and latent assumptions that would otherwise remain unarticulated. What is at stake is not only which values are present, but how they are framed—whether, for example, efficiency is understood purely in energetic terms or in relation to broader sustainability considerations [24-29].

Once these values are articulated, attention shifts toward the distribution of their relevance across different actors and systems. Materials AI operates within a heterogeneous landscape that includes scientific communities, industrial stakeholders, regulatory institutions, end-users, and environmental systems, each of which may be affected in distinct ways. Mapping these relationships clarifies whose interests are implicated and reveals asymmetries in influence that may otherwise go unexamined. In practice, such mapping exposes the ways in which certain priorities—often those aligned with technical performance or economic return—become structurally privileged over others.

Translating identified values into actionable design criteria constitutes a critical inflection point within the framework. Abstract commitments must be reformulated as measurable constraints or objectives that can be embedded within computational workflows. Sustainability, for instance, may be expressed through constraints on embodied carbon or incorporated as penalty terms within loss functions, while transparency may necessitate the integration of interpretable model components. This translation does not merely formalize values; it reshapes the optimization landscape itself by redefining what counts as a successful outcome.

The incorporation of these operationalized values into the technical pipeline ensures that they exert influence at the level where decisions are materially enacted. Data selection protocols, model architectures, objective functions, and evaluation metrics are all adjusted to reflect the identified priorities, allowing values to guide system behavior from within rather than being appended after optimization has already occurred. Under these conditions, the distinction between technical and normative design begins to dissolve, as both become co-constitutive elements of the same process.

Evaluation then assumes a dual role, extending beyond conventional performance metrics to include assessment of how effectively the system aligns with the values it was designed to embody. This requires the integration of both quantitative indicators and qualitative feedback from affected stakeholders, enabling a more comprehensive understanding of system behavior. Importantly, evaluation is not treated as a terminal stage but as a mechanism for iterative refinement, feeding back into earlier stages of identification, mapping, and translation. Through this recursive structure, the framework resists linearity and instead supports continuous adjustment, ensuring that value alignment remains an ongoing and adaptive process rather than a one-time design decision [21-28].

Figure 1 presents the hierarchical architecture of value-sensitive materials AI, showing how value identification is progressively translated into stakeholder analysis, measurable design requirements, technical implementation, and evaluative assessment.

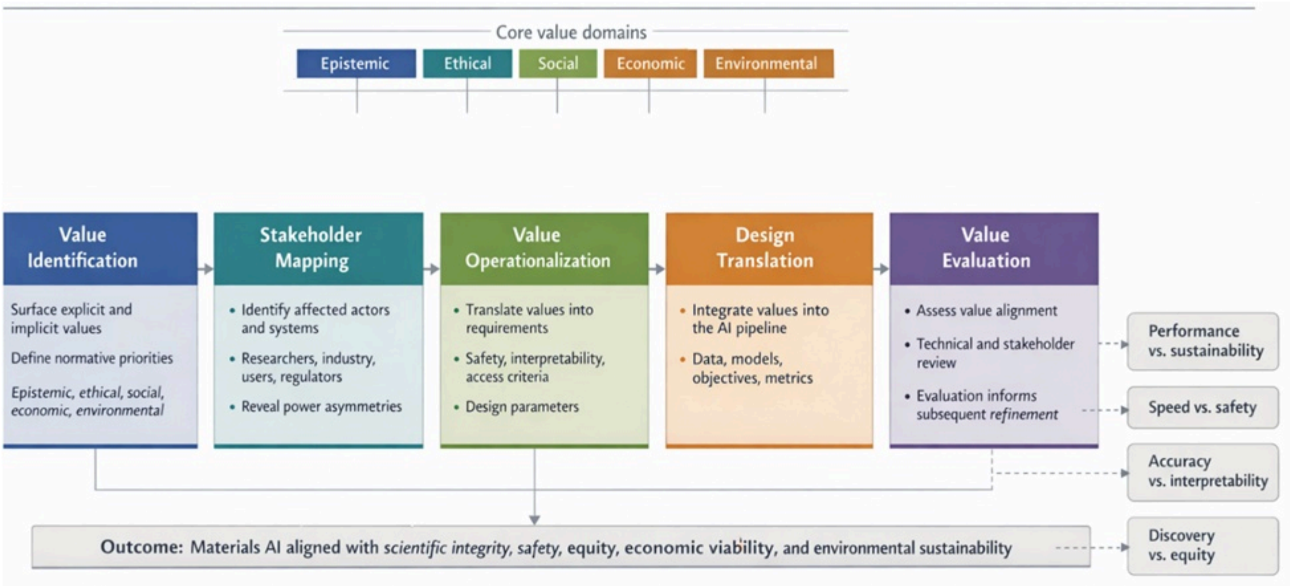

Figure 1. The hierarchical architecture of value-sensitive materials AI.

Types of Values in Materials AI

A typology of values relevant to materials AI systems provides a structured way to recognize the multiplicity of normative commitments that shape research and deployment. This paper articulates five distinct but interconnected value categories—epistemic, ethical, social, economic, and environmental—each of which can be identified, operationalized, and evaluated within the five-component framework proposed above.

Table 1 translates the five major value categories into concrete pipeline locations, technical design levers, and evaluative indicators, thereby clarifying how abstract normative commitments can become operational within materials AI.

Table 1. Translating value categories into materials AI design parameters, technical levers, and evaluative indicators

Value category	Core concern in materials AI	Typical locus in the pipeline	Technical design levers	Example evaluative indicators

Epistemic	Reliability and scientific validity of knowledge claims	Data curation, feature selection, model selection, validation	Uncertainty quantification, interpretable models, cross-domain validation, error calibration, and reproducibility protocols	Out-of-sample robustness, uncertainty calibration, reproducibility across datasets/labs, and mechanistic plausibility
Ethical	Prevention of harm and enforceable responsibility	Screening, optimization constraints, and deployment decisions	Toxicity filters, safety constraints, audit trails, explainability layers, and dual-use screening rules	Safety threshold compliance, traceability of recommendations, explainability adequacy, and documented accountability
Social	Distribution of benefits and burdens across groups and regions	Problem framing, application targeting, access conditions, and governance design	Inclusive stakeholder input, open-data/open-model options, public-interest application criteria, and benefit-sharing provisions	Accessibility of outputs, inclusiveness of stakeholder representation, public-benefit orientation, and equity of downstream access
Economic	Feasibility and real-world viability of discovered materials	Objective functions, manufacturability screening, and scale-up decisions	Cost-aware optimization, manufacturability constraints, supply-chain feasibility checks, and techno-economic integration	Estimated production cost, scalability potential, supply stability, and infrastructure compatibility
Environmental	Ecological sustainability across the material life cycle	Candidate screening, objective design, and post-discovery selection	Life-cycle proxies, embodied carbon penalties, recyclability criteria, non-toxicity constraints, and circularity-aware objectives	Carbon footprint, recyclability potential, toxicity risk, resource criticality, and end-of-life recovery potential

Type 1: Epistemic values

Epistemic values concern the quality and reliability of knowledge produced by materials AI systems. They include accuracy, reproducibility, interpretability, and scientific validity. In materials AI, epistemic values are enacted whenever a model is trained to predict properties such as formation energy or mechanical strength; the choice of loss function or cross-validation protocol implicitly privileges certain forms of reliability over others. For example, high predictive accuracy on benchmark datasets drawn from high-throughput computations may come at the expense of generalizability to experimentally synthesized materials, thereby undermining scientific validity. Operationalization might involve incorporating uncertainty quantification modules or requiring models to report confidence intervals alongside predictions, ensuring that epistemic humility is built into the system rather than treated as an afterthought.

Type 2: Ethical values

Ethical values focus on preventing harm and upholding moral responsibilities. Key examples are safety, fairness, accountability, and transparency. Materials AI systems can embed ethical values by screening for toxic or unstable compounds before suggesting them for synthesis, or by designing algorithms that flag potential dual-use risks in energetic materials. Fairness arises when datasets disproportionately represent materials from well-resourced laboratories, potentially biasing discovery toward applications relevant only to wealthy nations. Transparency requires that the reasoning behind a recommended material be explainable to non-expert stakeholders, such as regulators or community groups. These values are operationalized through constraint layers in generative models or post-hoc explanation techniques that link predictions to underlying physical mechanisms.

Type 3: Social values

Social values address questions of justice, access, and collective benefit. They encompass equity, public good, benefit sharing, and inclusivity. In the materials AI context, social values are at stake when new battery materials discovered through autonomous workflows are patented exclusively by large corporations, limiting access for researchers in the Global South. Equity demands that discovery pipelines consider applications relevant to underserved communities, such as low-cost water purification materials or affordable housing composites. Benefit sharing might be operationalized by requiring open-access publication of promising leads or by designing licensing frameworks that prioritize public research institutions. Stakeholder mapping within the framework ensures that voices from civil society and developing regions inform value identification from the outset.

Type 4: Economic values

Economic values emphasize resource optimization and viability in real-world deployment. They include efficiency, cost-effectiveness, scalability, and return on investment. Materials AI frequently prioritizes these values when objective functions minimize synthesis cost or maximize yield under industrial constraints. For instance, a model trained to accelerate the discovery of photovoltaic materials may implicitly favor solutions that scale rapidly in existing manufacturing infrastructure, even if they require rare elements. Operationalization can occur through multi-objective optimization that incorporates techno-economic analysis as a core term rather than an external filter applied after discovery. This integration ensures economic considerations shape the search space itself rather than merely pruning it.

Type 5: Environmental values

Environmental values center on planetary health and long-term ecological integrity. Prominent among them are sustainability, non-toxicity, circularity, and reduced carbon footprint. Materials informatics pipelines that optimize solely for performance metrics risk overlooking the environmental legacy of proposed compounds, such as persistent pollutants or high embodied energy. Sustainability can be embedded by adding life-cycle assessment proxies directly into surrogate models, while circularity might favor materials designed for recyclability or biodegradability. These values gain particular urgency in light of growing recognition that materials discovery must align with global sustainability goals.

Each type is not isolated; they interact dynamically. An epistemic commitment to interpretability may enhance ethical transparency and social trust, while environmental values often stand in tension with short-term economic efficiency. The typology, therefore, serves as a diagnostic tool within the value identification component, enabling researchers to surface hidden assumptions and to negotiate priorities deliberately rather than allowing default technical heuristics to dictate outcomes. By making these five categories explicit, the framework transforms materials AI from a value-implicit endeavor into one that systematically accounts for the full spectrum of stakes involved.

Value Tensions and Trade-Offs

Value tensions arise inevitably when multiple value categories pull design decisions in opposing directions. Materials AI systems frequently encounter four archetypal tensions that the proposed framework is designed to surface and navigate rather than suppress.

Table 2 consolidates the manuscript’s core trade-off logic by specifying how recurrent value tensions in materials AI map onto stakeholder exposure, technical resolution strategies, and governance questions requiring explicit justification.

Table 2. Archetypal value tensions in materials AI: conflict structure, affected stakeholders, and design-resolution strategies

Value tension	What is in conflict	Typical manifestation in materials AI	Stakeholders most affected	Design-resolution strategies within VSD	Key governance question
Performance vs. sustainability	Maximizing target properties vs. minimizing ecological burden	High-performing candidates depend on scarce, toxic, or energy-intensive inputs	Environmental regulators, affected communities, future generations, and industry	Multi-objective optimization, embodied-carbon constraints, resource-criticality penalties, and life-cycle-informed screening	What level of performance gain justifies additional environmental cost?
Speed vs. safety	Rapid discovery cycles vs. rigorous hazard evaluation	Fast autonomous screening advances candidates before sufficient toxicity or stability assessment	Regulators, laboratory workers, end-users, and exposed communities	Mandatory safety gates, staged validation protocols, high-risk candidate flagging, and documented threshold rules	Which safety checks are non-negotiable before progression to synthesis or deployment?
Accuracy vs. interpretability	Predictive performance vs. transparency and scientific understanding	Black-box models outperform simpler models but provide a weak mechanistic explanation	Scientists, reviewers, regulators, downstream adopters	Hybrid models, post-hoc explanation tools, interpretable surrogate models, and explanation reporting standards	When is lower explainability acceptable, and under what accountability conditions?
Discovery vs. equity	Concentration of innovation gains vs. fair distribution of benefits	Valuable discoveries are patented or deployed in ways that exclude underserved regions or the public	Public institutions, low-resource communities, Global South researchers, and civil society	Stakeholder-inclusive problem selection, open science pathways, public-interest licensing, and access-sensitive evaluation criteria	Who benefits from discovery, and who is excluded from access or exposed to risk?

Performance vs. sustainability

High-performance metrics—such as superior thermal stability or energy density—often conflict with environmental values when candidate materials require energy-intensive synthesis or contain scarce or toxic elements. A generative model optimized exclusively for property prediction may converge on solutions whose life-cycle impacts are unacceptable. The framework addresses this through value operationalization and design translation: sustainability constraints are encoded as hard penalties or multi-objective Pareto fronts, allowing designers to explore trade-off surfaces explicitly rather than discovering them post hoc.

Speed vs. safety

Accelerated discovery workflows prized in autonomous laboratories can bypass rigorous safety screening, prioritizing rapid iteration over thorough toxicity or stability assessment. This tension is exacerbated when computational budgets favor quick surrogate models over more computationally expensive but safer physics-informed approaches. Stakeholder mapping reveals how industry partners may emphasize speed while regulatory bodies insist on safety; the iterative value evaluation component then requires documented justification whenever safety thresholds are relaxed.

Accuracy vs. interpretability

Black-box deep learning architectures frequently achieve superior predictive accuracy yet sacrifice the epistemic and ethical value of interpretability. In materials contexts, an opaque model might correctly identify a promising alloy but offer no mechanistic insight, hindering both scientific understanding and regulatory approval. The framework resolves this by embedding interpretability requirements at the design translation stage—favoring hybrid architectures or post-hoc explanation layers—while still allowing accuracy to be traded off transparently against other values.

Discovery vs. equity

Rapid discovery of novel materials can concentrate benefits among well-funded institutions or commercial entities, leaving marginalized communities bearing disproportionate risks or excluded from access. For example, AI-driven development of advanced composites for aerospace may advance economic competitiveness while exacerbating global material supply-chain inequities. Value-sensitive stakeholder mapping and operationalization ensure that equity considerations—such as open-data mandates or targeted application scenarios—are integrated into objective functions, turning potential trade-offs into deliberate, negotiated design parameters.

In each case, the five-component framework does not eliminate tensions but renders them visible and manageable. By cycling through identification, mapping, operationalization, translation, and evaluation, researchers can document the reasoning behind chosen trade-offs, making normative commitments traceable and contestable. This structured navigation distinguishes value-sensitive materials AI from approaches that treat value conflicts as mere engineering optimizations.

Objections and Replies

Several objections are frequently raised against the integration of explicit value analysis in materials AI. Yet, closer examination reveals that they rest on a mischaracterization of how technical systems are actually constructed. A recurring concern is that the introduction of values undermines scientific objectivity, on the assumption that normative considerations are inherently subjective and therefore incompatible with rigorous inquiry. This position, however, overlooks the extent to which materials AI systems already encode value-laden assumptions through routine technical decisions. Choices concerning data inclusion, objective formulation, and evaluation criteria are never neutral; they implicitly privilege certain outcomes, representations, and forms of success over others. Making these commitments explicit does not compromise objectivity but strengthens it by replacing implicit, unexamined defaults with transparent and contestable reasoning. In this sense, value articulation functions as a mechanism of epistemic clarification rather than distortion.

A related critique attempts to separate value considerations from core technical work by relegating them to the domain of ethics, as though they were external to the design of algorithms and models. Such a division proves untenable in practice. The architecture of a model, the structure of its training data, and the metrics used to evaluate its performance are all shaped by prior assumptions about what matters and why. Treating these assumptions as irrelevant to technical research does not preserve neutrality; it obscures the normative dimensions that already guide system behavior. From this perspective, the exclusion of value analysis is not a form of disciplinary rigor but a limitation that constrains the alignment between system design and intended outcomes. Incorporating value-sensitive considerations, by contrast, enhances technical coherence by ensuring that optimization targets reflect the broader purposes for which the system is developed.

Another line of resistance suggests that materials AI remains at too early a stage of development for such concerns to be meaningful, proposing that normative reflection should follow rather than accompany technical maturation. This argument assumes that values can be retrofitted once systems are fully developed, yet the structure of design processes in materials AI suggests otherwise. Early decisions—such as which datasets are assembled, which properties are prioritized, and which modeling paradigms are adopted—establish trajectories that are difficult to alter retrospectively. These initial commitments shape the evolution of the system, embedding path dependencies that constrain future possibilities. Under these conditions, postponing value analysis does not defer its influence; it simply allows it to operate implicitly and without scrutiny. Engaging with values at early stages, therefore, represents not a premature intervention but a strategic necessity, ensuring that systems remain adaptable and aligned with evolving scientific and societal priorities.

Taken together, these responses indicate that resistance to value-sensitive approaches often stems from a conceptual separation between technical practice and normative reasoning that does not hold under closer analysis. By demonstrating how values are already embedded within the fabric of materials AI, the framework reframes their explicit consideration as an extension of technical rigor rather than a departure from it, offering practical tools for integrating reflection and design in a mutually reinforcing manner.

Implications for Materials AI Practice

Adopting the proposed value-sensitive framework carries concrete implications for authors, reviewers, and the broader community.

For authors, three changes are essential. First, manuscripts should explicitly state the values guiding design choices, linking them to specific components of the framework. Second, authors must report stakeholder considerations and any value trade-offs negotiated during model development. Third, discussion sections should analyze how the chosen technical implementation supports or compromises identified values rather than focusing solely on performance metrics.

For reviewers, evaluation criteria should expand beyond methodological soundness and benchmark scores. Reviewers should ask whether value assumptions have been articulated, whether stakeholder mapping was performed, and whether operationalization choices are justified. Papers claiming “value-neutral” or purely technical contributions should be scrutinized for implicit normative commitments that remain unexamined.

For the community as a whole, three institutional developments are needed. First, value reporting standards—analogous to existing data and code availability requirements—should be developed and adopted by leading journals. Second, training materials and curricula on value-sensitive design must be integrated into materials science and AI graduate programs. Third, funding agencies should prioritize research that advances both technical frontiers and value alignment, creating incentives that align with the framework’s goals.

These shifts do not diminish the technical ambitions of materials AI; they enhance them by ensuring that powerful discovery tools serve well-articulated human and environmental ends.

Conclusion

This paper has argued that materials AI systems inevitably embed values despite frequent claims of neutrality. It has defined values in the materials AI context, diagnosed four reasons for their neglect, introduced the origins and principles of Value-Sensitive Design, and proposed a five-component framework—value identification, stakeholder mapping, value operationalization, design translation, and value evaluation—tailored to the unique demands of materials discovery. A typology of five value categories (epistemic, ethical, social, economic, and environmental) together with an analysis of key tensions and responses to common objections completes the conceptual structure.

The framework offers researchers a practical yet philosophically grounded pathway to move from implicit value embedding to deliberate value alignment. Its cyclical design ensures that value sensitivity becomes an ongoing reflexive practice rather than a one-time compliance exercise. As materials AI continues to reshape how new substances are imagined and brought into existence, the adoption of value-sensitive practices must become a standard component of responsible research. Only by making values visible, negotiable, and actionable can the field ensure that its powerful tools serve not merely technical metrics but the broader imperatives of sustainability, equity, safety, and scientific integrity. The future of materials AI lies in systems that are both technically excellent and normatively reflective—systems that accelerate discovery while safeguarding the planetary and social commons upon which all innovation ultimately depends.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 12 Aug 2022 Revised: 27 Sep 2022 Accepted: 08 Dec 2022
Published online: 18 January 2023

References

Islind AS, Willermark SM. Becoming a designer: The value of sensitive design situations for teaching and learning ethical design and design theory. Scand J Inf Syst. 2022;34(1):1.

Friedman B, Hendry DG. Value sensitive design: Shaping technology with moral imagination. Cambridge: MIT Press; 2019. https://doi.org/10.7551/mitpress/7585.001.0001.

Van Der Hoven J, Manders-Huits N. Value-sensitive design. In: The ethics of information technologies. London: Routledge; 2020. p. 329-32.

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).

Umbrello S, Van de Poel I. Mapping value sensitive design onto AI for social good principles. AI Ethics. 2021;1(3):283-96.

Oviedo F, Ferres JL, Buonassisi T, Butler KT. Interpretable and explainable machine learning for materials science and chemistry. Acc Mater Res. 2022;3(6):597-607.

Melia HR, Muckley ES, Saal JE. Materials informatics and sustainability—the case for urgency. Data Cent Eng. 2021;2:e19.

Pilania G. Machine learning in materials science: From explainable predictions to autonomous design. Comput Mater Sci. 2021;193:110360.

Hong Y, Hou B, Jiang H, Zhang J. Machine learning and artificial neural network accelerated computational discoveries in materials science. Wiley Interdiscip Rev Comput Mol Sci. 2020;10(3):e1450.

López C. Artificial intelligence and advanced materials. Adv Mater. 2023;35(23):2208683.

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

Allen AE, Tkatchenko A. Machine learning of material properties: Predictive and interpretable multilinear models. Sci Adv. 2022;8(18):eabm7185.

Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A, Kim C. Machine learning in materials informatics: Recent applications and prospects. npj Comput Mater. 2017;3(1):54.

Umbrello S. Beneficial artificial intelligence coordination by means of a value sensitive design approach. Big Data Cogn Comput. 2019;3(1):5.

Choudhary K, DeCost B, Chen C, Jain A, Tavazza F, Cohn R, et al. Recent advances and applications of deep learning methods in materials science. npj Comput Mater. 2022;8(1):59.

Nasteski V. An overview of the supervised machine learning methods. Horizons. 2017;4:51-62.

Piovesan D, Tan JB, Becker A, Banuelos J, Narasappa N, DiRenzo D, et al. Targeting CD73 with AB680 (Quemliclustat), a novel and potent small-molecule CD73 inhibitor, restores immune functionality and facilitates antitumor immunity. Mol Cancer Ther. 2022;21(6):948-59.

Friedman B, Hendry DG, Borning A. A survey of value sensitive design methods. Found Trends Hum Comput Interact. 2017;11(2):63-125.

Umbrello S, De Bellis AF. A value-sensitive design approach to intelligent agents. In: Artificial intelligence safety and security. Boca Raton: CRC Press; 2018. p. 395-409.

Stein HS, Gregoire JM. Progress and prospects for accelerating materials science with automated and autonomous workflows. Chem Sci. 2019;10(42):9640-9.

Suh C, Fare C, Warren JA, Pyzer-Knapp EO. Evolving the materials genome: How machine learning is fueling the next generation of materials discovery. Annu Rev Mater Res. 2020;50(1):1-25.

National Academies of Sciences, Engineering, and Medicine. Autonomous materials discovery and optimization: Proceedings of a workshop–in brief. Washington DC: National Academies Press; 2021.

Ting JM, Tamayo-Mendoza T, Petersen SR, Van Reet J, Ahmed UA, Snell NJ, et al. Frontiers in nonviral delivery of small molecule and genetic drugs, driven by polymer chemistry and machine learning for materials informatics. Chem Commun. 2023;59(96):14197-209.

Brown DS, Schneider J, Dragan A, Niekum S. Value alignment verification. In: International Conference on Machine Learning. Brookline: PMLR; 2021. p. 1105-15.

Lv Z, Poiesi F, Dong Q, Lloret J, Song H. Deep learning for intelligent human–computer interaction. Appl Sci. 2022;12(22):11457.

Boyd K. Designing up with value-sensitive design: Building a field guide for ethical ML development. In: Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. New York: ACM; 2022. p. 2069-82.