ORIGINAL RESEARCH **Open access**

The Problem of Normative Assumptions in Materials AI Objective Functions

Ricardo Alves[1], Bruno Costa[1*], Helena Martins[2], Nuno Faria[1]

Abstract

Objective functions in materials artificial intelligence are routinely presented as neutral computational devices that merely minimize formation energy or maximize ionic conductivity. Yet, they quietly embed normative assumptions about what constitutes a “good” material, thereby encoding ethical, social, and scientific value judgments that shape downstream discovery pathways. These functions operate as value vehicles by translating ostensibly descriptive metrics into prescriptive targets that privilege certain outcomes—stability over metastability, efficiency over sustainability—while rendering competing priorities invisible. This critique identifies four interlocking problems: the hidden normativity concealed within technical loss functions, the resulting value monoculture that narrows the space of desirable materials, the measurability trap that biases discovery toward easily quantified properties, and the democratic deficit that excludes affected stakeholders from objective formulation. The consequences of these unexamined assumptions include narrow and path-dependent discovery trajectories, unresolved value conflicts, and a systematic exclusion of societal considerations from materials innovation. Alternative approaches are therefore proposed that treat objective design as an explicit exercise in value articulation, multi-objective negotiation, and participatory governance, thereby transforming materials AI from a value-blind optimizer into a reflexive, value-aware sociotechnical practice.

Keywords Materials AI, Objective functions, Normative assumptions, Value-laden optimization, Hidden normativity, Value monoculture

*Correspondence:
Bruno Costa
bruno.costa@gmail.com
[1] Department of AI Materials Engineering, University of Porto, Porto, Portugal
[2] Department of Computational Materials Science, University of Aveiro, Aveiro, Portugal

Introduction

The problem confronting contemporary materials artificial intelligence is not merely technical but profoundly philosophical: objective functions that drive inverse design, generative modeling, and property optimization are treated as purely technical choices—minimize formation energy, maximize ionic conductivity, achieve a target bandgap—yet these choices encode normative judgments about what is good, valuable, or desirable in a material [1-5]. Far from being value-neutral instruments, such functions instantiate decisions about which properties matter, which trade-offs are acceptable, and whose interests should be prioritized, thereby collapsing the classical distinction between descriptive and normative objectives. A descriptive objective, such as accurately predicting a formation energy from first-principles data, seeks only to represent what is; a normative objective, by contrast, asserts that a particular formation energy ought to be minimized because stability is deemed inherently superior to metastability or because computational tractability outweighs long-term environmental cost [1, 6]. Materials AI has largely ignored this entanglement, proceeding as if optimization targets exist in a vacuum of facts rather than a field of values.

This paper advances a critical critique grounded in the philosophy of science, ethics of technology, and the specific literature on machine learning for materials [7, 8]. It argues that the hidden normativity of objective functions is not an accidental oversight but a structural feature of current practice. By examining how common objectives in materials discovery embed unacknowledged values, the analysis reveals a value monoculture that privileges efficiency, performance, and novelty at the expense of sustainability, safety, equity, and accessibility [9, 10]. Four interlocking critique points are developed: the concealment of normativity behind technical language, the narrowing of value horizons into a monoculture, the measurability trap that substitutes computable proxies for what truly matters, and the absence of democratic input into decisions that affect global material futures.

The critique proceeds by first establishing objective functions as vehicles of value, then dissecting five canonical materials AI objectives to expose their implicit normative commitments. Subsequent sections elaborate on each of the four critique points in turn, trace the downstream consequences for materials discovery, and outline constructive alternatives that render value choices explicit and contestable. Throughout, the argument draws exclusively upon the peer-reviewed corpus identified in the reference compilation, treating Putnam's collapse of the fact/value dichotomy and Russell's call for human-compatible AI as foundational touchstones while integrating recent contributions from machine-learning-for-materials and science-and-technology-studies literatures [1-3]. The ultimate aim is not to reject optimization but to insist that materials AI become reflexively aware of the values it already encodes, thereby opening the possibility of more plural, accountable, and democratically legitimate discovery pathways.

The stakes extend beyond academic debate. Materials discovered through value-laden objectives will populate the built environment, energy systems, and biomedical devices of the coming decades; the normative assumptions locked into today's loss functions will therefore co-produce tomorrow's sociotechnical realities [11, 12]. If these assumptions remain unexamined, materials AI risks reproducing and amplifying the very inequities and environmental externalities it claims to solve through technological progress. This introduction, therefore, frames the remainder of the manuscript as both diagnosis and prescription: a sustained argument that objective functions are never merely technical and that their design demands the same critical scrutiny applied to any other value-laden scientific practice [13, 14].

Objective Functions as Value Vehicles

Objective functions in materials AI are not neutral technical artifacts; they are vehicles that transport normative commitments from the mind of the designer into the behavior of the algorithm [1]. Putnam argued that the fact/value dichotomy collapses once one recognizes that even the most seemingly descriptive statements presuppose evaluative frameworks [1], and objective functions exemplify this entanglement with particular clarity. When a researcher defines a loss function as the mean absolute error between predicted and DFT-computed formation energies, the choice appears descriptive—merely a measure of predictive fidelity—yet it simultaneously enacts a normative judgment that predictive accuracy on energy is the paramount good against which all other modeling choices must be measured [3, 15].

The value-carrying capacity of objective functions operates through three interlocking mechanisms. First, they define what counts as "good" or "optimal" by assigning scalar rewards or penalties to specific property combinations; in doing so, they legislate a hierarchy of desirability among possible materials. Second, they prioritize certain outcomes over others by weighting terms in multi-objective formulations or by selecting single-objective simplifications that render competing values invisible. Third, they embed assumptions about acceptable trade-offs—computational cost versus physical fidelity, short-term performance versus long-term sustainability—without ever rendering those assumptions explicit or contestable [2, 16]. Russell notes that value alignment is central to AI safety [2], yet materials AI has largely ignored this insight, treating alignment as an afterthought rather than a foundational design constraint.

One can conceptualize the value embedding process as a pipeline in which raw technical parameters—atomic coordinates, electronic densities, phonon spectra—enter an objective function that applies a series of selective filters. Each filter corresponds to an implicit normative stance: lower energy is better, faster ion transport is better, and larger bandgaps for certain applications are better. These filters do not merely describe reality; they prescribe which slices of

reality deserve attention and resources. The pipeline, therefore, transforms an ostensibly descriptive computational workflow into a normative engine that steers discovery toward materials that satisfy the embedded values while discarding those that do not [17, 18].

The distinction between descriptive and normative objectives becomes especially salient here. A purely descriptive model might seek to reproduce experimental observables with minimal error; the corresponding objective function would be evaluated solely on its representational fidelity. Once that model is embedded in an optimization loop, however, the same error metric acquires normative force: the algorithm is now directed to search for materials that minimize the very quantity the model was trained to predict accurately. The transition from description to prescription is seamless and rarely acknowledged, yet it is precisely where value enters the system [1, 19].

Literature in philosophy of science and ethics of information technology has long recognized that optimization targets are never innocent [20, 21]. In materials AI, however, this recognition has been slow to arrive. Butler *et al.* [4] survey machine learning for molecular and materials science without once addressing the normative loading of the objectives they review. At the same time, Schmidt *et al.* [5] celebrate recent advances in solid-state applications while treating target properties as self-evident goods. Such omissions are not mere oversights; they reflect a disciplinary culture that regards objective functions as technical parameters rather than sites of ethical and political deliberation [22, 23].

By contrast, the present critique insists that every objective function is a value vehicle. It carries assumptions about human needs (energy storage over energy equity), environmental priorities (performance over planetary boundaries), and epistemic virtues (computational elegance over epistemic humility). Recognizing this fact is the first step toward responsible design. Until materials scientists treat objective formulation as an explicit exercise in value articulation, the field will continue to reproduce hidden normativities that constrain rather than liberate discovery [3, 24].

Figure 1 shows that objective functions in materials AI operate not as neutral mathematical devices but as normative architectures that encode value commitments, generate systematic distortions in discovery, and therefore require reflexive governance.

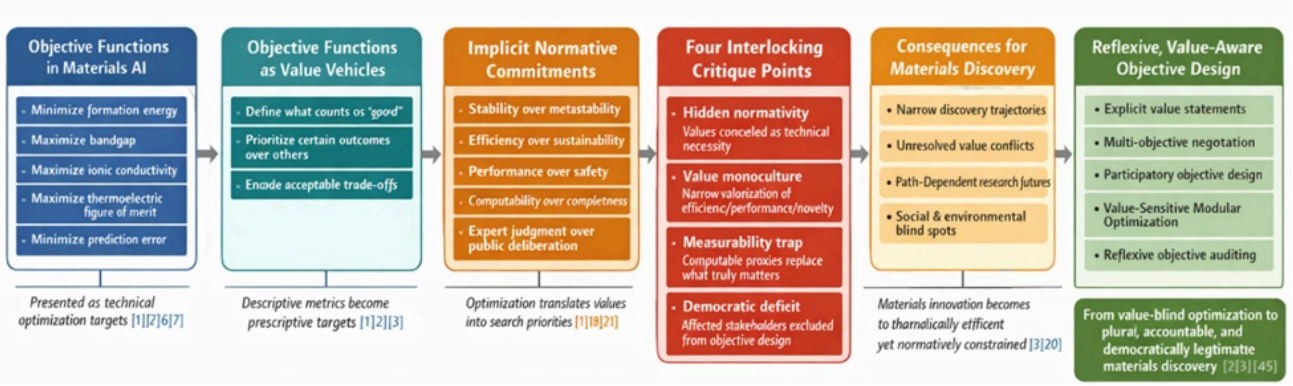

Figure 1. Normative architecture of objective functions in materials AI: From hidden value encoding to reflexive governance.

Common Materials AI Objectives

Materials AI routinely optimizes a narrow repertoire of objectives, each of which appears technically straightforward yet conceals a cluster of normative assumptions. Five canonical examples suffice to illustrate the pattern.

Minimize formation energy

The objective is ubiquitous in generative models and inverse design pipelines because formation energy is readily computed via density-functional theory and serves as a proxy for thermodynamic stability [4, 7]. The hidden normative commitment is that stability is inherently preferable to metastability, that ground-state structures are more “natural” or “valuable” than higher-energy polymorphs. That computational convenience justifies privileging one axis of material performance over others, such as synthesizability under non-equilibrium conditions or long-term kinetic persistence [6, 25].

Maximize bandgap

In optoelectronic and photovoltaic applications, algorithms frequently seek to enlarge the electronic bandgap to a target window [5, 26]. The normative loading here privileges insulation or specific electronic transport characteristics over metallic conductivity or topological protection, implicitly endorsing a vision of materials utility centered on solar-energy conversion or transistor performance while marginalizing applications that might require semimetallic or superconducting states. The choice further assumes that bandgap magnitude is the decisive figure of merit rather than, for example, defect tolerance or environmental stability [27].

Maximize ionic conductivity

Solid-electrolyte and battery research routinely optimizes for high room-temperature ionic conductivity [4, 28]. The value judgment embedded in this target is that rapid ion transport and, therefore, high-rate energy storage constitute a preeminent societal good. Less visible are the trade-offs with electrochemical stability, mechanical robustness, or the use of critical raw materials; the objective silently endorses efficiency and performance while devaluing safety, resource scarcity, and circular-economy considerations [29].

Maximize thermoelectric figure of merit (zT)

Thermoelectric discovery pipelines target high zT by simultaneously optimizing Seebeck coefficient, electrical conductivity, and thermal resistivity [5]. The normative stance is that waste-heat recovery and energy-conversion efficiency are paramount, thereby prioritizing a narrow decarbonization pathway while backgrounding questions of material toxicity, scalability, or end-of-life recyclability.

Minimize prediction error

Across surrogate models and active-learning loops, the objective is often framed as minimizing mean-squared error or similar loss functions between model predictions and reference data [7]. At first glance, purely descriptive, this target becomes normative once embedded in optimization: it privileges models that reproduce existing data distributions, thereby reinforcing path dependence on historically measured properties and discouraging exploration of outlier regions where novel phenomena might reside. Accuracy is elevated to the status of an intrinsic good, even when that accuracy concerns properties that are themselves value-laden [3].

In each case, the objective function does more than guide search; it enacts a vision of what materials science should value. The literature routinely presents these targets as self-evident technical necessities [4-6], yet a critical reading reveals them as carriers of unacknowledged normative commitments.

Table 1 demonstrates that the most common objective functions in materials AI each carry a distinct normative signature that privileges some forms of value while systematically marginalizing others.

Table 1. Canonical materials, AI objective functions, and their embedded normative commitments

Canonical objective	Apparent technical	Hidden normative	Value privileged	Value marginalized	Typical epistemic effect on

function	purpose	assumption			discovery
Minimize formation energy	Favor thermodynamic stability and screen candidate materials efficiently	Ground-state stability is treated as inherently superior to metastability or kinetic persistence	Stability, tractability, formal neatness	Metastability, synthesizability under non-equilibrium conditions, and exploratory novelty	Biases search toward equilibrium structures and away from functionally valuable metastable candidates
Maximize bandgap	Target desired electronic structure windows for optoelectronic applications	A specific bandgap regime is assumed to define material usefulness	Device-oriented performance, application fit	Alternative electronic states, defect tolerance, and environmental durability	Narrow the discovery to pre-selected utility models rather than broader functional possibilities
Maximize ionic conductivity	Improve fast ion transport for batteries and solid electrolytes	High-rate transport is treated as the overriding criterion of material desirability	Efficiency, speed, and energy-storage performance	Safety, interfacial stability, raw-material ethics, and circularity	Produces candidates optimized for transport metrics while overlooking deployment-relevant harms and constraints
Maximize thermoelectric figure of merit (zT)	Enhance waste-heat conversion performance	Energy-conversion efficiency is assumed to outweigh toxicity, scalability, or end-of-life concerns	Efficiency, decarbonization performance, benchmark competitiveness	Recyclability, abundance, toxicity, and infrastructure compatibility	Encourages technically impressive but materially narrow pathways that may not scale responsibly
Minimize prediction error	Improve surrogate accuracy and guide active learning or inverse design	Fidelity to available data is treated as the primary good	Accuracy, benchmark success, and reproducibility within known distributions	Outlier exploration, epistemic humility, and novelty beyond the dataset	Reinforces path dependence by privileging already measured properties and historically dominant data regimes

Critique Point 1: Hidden Normativity

The first critique point concerns the systematic concealment of normativity behind the language of technical necessity. Objective functions appear as neutral mathematical constructs—simple scalars to be minimized or maximized, yet they embed value judgments that are never rendered explicit [1, 3]. Putnam's demonstration that facts and values are

inextricably entangled finds direct application here: the “fact” of a lower formation energy is immediately converted into the “value” that such a material is preferable, without any intervening justification [1].

Hidden normativity manifests most clearly in the rhetorical framing of papers that describe objective functions as “objective” in both the mathematical and epistemological senses. By labeling a loss function “objective,” authors imply that it stands outside the realm of human values, when in reality it is the very site where values are inscribed. Consider the widespread practice of framing formation-energy minimization as a straightforward thermodynamic imperative. The framing erases the prior normative decision that thermodynamic ground states deserve priority over kinetically stabilized or metastable phases that may exhibit superior functional properties under operating conditions [6, 25].

The concealment is not accidental. It serves to maintain the appearance of scientific neutrality, thereby shielding objective choices from ethical and political scrutiny [20]. When normativity is hidden, critique becomes difficult; one cannot contest a value that has been rendered invisible. The result is a disciplinary culture in which researchers optimize without acknowledging that they are simultaneously prescribing what counts as progress [2].

Russell’s insistence on explicit value alignment in advanced AI systems stands in stark contrast to current materials practice [2]. While general AI safety research has begun to grapple with the problem of misaligned objectives, materials AI continues to treat its loss functions as unproblematic technical parameters. The critique, therefore, calls for a reflexive turn: every objective function should be accompanied by an explicit value statement that articulates the normative commitments it enacts. Only then can the field move from hidden normativity to transparent value governance [3].

Critique Point 2: Value Monoculture

The second critique point diagnoses the emergence of a value monoculture within materials AI. Optimization targets converge on a narrow band of virtues—efficiency, performance, novelty, computational elegance—while systematically excluding competing values such as sustainability, safety, equity, accessibility, and long-term resilience [9, 10]. This monoculture is not the accidental byproduct of individual choices but the structural outcome of a disciplinary incentive structure that rewards papers reporting incremental improvements on canonical metrics.

Value monoculture operates through two reinforcing mechanisms. First, the community converges on a small set of benchmark objectives because they are computationally tractable and appear in high-impact publications; subsequent work then inherits and reinforces the same targets, creating a self-reinforcing loop [4, 5]. Second, multi-objective formulations, when they appear, typically trade off only within the accepted value set—e.g., energy density versus power density—rather than across incommensurable societal values such as performance versus environmental justice.

The consequences are visible across application domains. Optimization for battery ionic conductivity routinely yields materials that rely on cobalt or lithium extracted under ethically problematic conditions. Yet, the objective function itself contains no term that penalizes supply-chain harm [28]. Similarly, thermoelectric materials optimized for peak zT frequently incorporate rare or toxic elements without any countervailing sustainability constraint. The monoculture, therefore, does not merely omit alternative values; it actively suppresses them by rendering them computationally invisible.

A pluralist alternative would require explicit incorporation of heterogeneous value dimensions into the objective function, yet current practice treats value pluralism as methodologically inconvenient rather than epistemically necessary [2]. By maintaining a monoculture of efficiency and performance, materials AI risks producing technically impressive yet socially and environmentally maladaptive materials. The critique, therefore, demands a deliberate broadening of the value horizon: objective functions must be designed to accommodate, rather than erase, the full spectrum of societal priorities that materials science claims to serve [3].

Critique Point 3: The Measurability Trap Revisited

The third critique point exposes what this analysis terms the measurability trap: materials AI inevitably optimizes what can be readily quantified within existing computational frameworks rather than what ultimately matters for real-world performance, societal benefit, or long-term viability. Formation energy, for instance, is eminently computable through density-functional theory or surrogate models. Yet, long-term kinetic stability under operating conditions, environmental degradation pathways, or supply-chain externalities remains stubbornly difficult to encode in differentiable loss functions [4, 5]. The trap arises because the optimization machinery demands gradients and scalar rewards; therefore, only those properties that admit efficient forward simulation or cheap oracle evaluation are elevated to the status of primary objectives. Everything else—toxicity under cyclic loading, end-of-life recyclability, or equity implications of raw-material sourcing—is either ignored or relegated to post-hoc filtering, which itself becomes another value-laden gatekeeping step.

This bias is not a minor technical limitation but a profound epistemological distortion. Schmidt and colleagues celebrate the rapid expansion of machine-learning applications in solid-state materials [5]. Yet, their survey implicitly accepts that the most successful pipelines are those built around properties already well-represented in high-throughput databases. The very success of such pipelines reinforces the trap: researchers gravitate toward objectives that yield publishable results quickly, further entrenching the preference for the measurable over the meaningful. Zunger's influential framework for inverse design similarly privileges target functionalities that can be expressed as explicit mathematical constraints [6], thereby systematically sidelining emergent phenomena whose value cannot be reduced to a single scalar. The result is a discovery landscape populated by materials that excel on paper (or in simulation) but frequently disappoint in deployment precisely because the unmeasured dimensions dominate real-world behavior.

A concrete illustration appears in battery-electrolyte design. Maximizing ionic conductivity is straightforward when the objective is framed as a room-temperature diffusion coefficient derived from molecular-dynamics snapshots or nudged-elastic-band calculations [7, 28]. The optimization proceeds smoothly, generating candidate solids with impressive simulated transport numbers. Yet the same materials may exhibit catastrophic interfacial instability or dendrite formation over hundreds of cycles—phenomena that are computationally expensive to simulate at scale and therefore rarely folded into the primary loss function. The algorithm has optimized the measurable proxy while the actual performance metric of interest (cycle life under realistic conditions) remains outside the optimization loop. Similarly, thermoelectric pipelines that maximize the figure of merit zT rely on easily computed electronic and thermal transport coefficients, but rarely incorporate full-lifecycle assessments of embodied carbon or mining impacts because those quantities resist differentiable formulation.

The measurability trap is compounded by the feedback loop between data availability and objective selection. High-throughput computational databases contain abundant formation energies and bandgaps because these quantities are cheap to compute; they contain far fewer entries for, say, fracture toughness under humid environments or biocompatibility under physiological pH [4]. Consequently, the surrogate models trained on such data inherit the same blind spots, and active-learning campaigns preferentially acquire more data on the already-well-measured axes. The trap, therefore, becomes self-perpetuating: measurability begets more measurement, which in turn deepens the bias. Russell's broader argument for human-compatible AI underscores the danger [2]; when objectives are chosen for computational convenience rather than human relevance, the resulting systems are aligned with the machine's capabilities rather than with societal needs.

Critically, the trap is not inevitable. It is an artifact of design choices that privilege tractability over completeness. Putnam's collapse of the fact/value dichotomy is again illuminating here [1]: the "fact" that a property is computable is immediately converted into the unstated "value" that it deserves primacy. The descriptive ease of calculating formation energy is treated as sufficient justification for its normative elevation, without ever interrogating whether computational accessibility aligns with the broader purposes of materials innovation. The present critique, therefore, insists that the measurability trap must be revisited and dismantled. Objective functions should be redesigned to incorporate proxy penalties or multi-fidelity hierarchies that explicitly acknowledge the limits of current measurability, thereby forcing the field to confront the gap

between what can be optimized and what should be optimized. Until that confrontation occurs, materials AI will continue to deliver elegant solutions to the wrong problems [3, 20].

The trap also carries ethical weight. By optimizing only the measurable, the field implicitly declares that unmeasurable harms—ecological disruption, community displacement, or future-generation burdens—do not count within the optimization calculus. This is not neutrality; it is a normative stance that privileges the visible and the immediate over the diffuse and the long-term [23]. A reflexive materials AI must therefore treat measurability itself as a value-laden parameter, subject to the same critical scrutiny applied to any other modeling choice. Only then can the discipline escape the trap and begin optimizing for properties that genuinely matter.

Critique Point 4: Who Decides?

The fourth critique point confronts the democratic deficit at the heart of objective-function design: the question of who is entitled to define what counts as a "good" material remains almost exclusively in the hands of a narrow technical elite—primarily academic researchers, national-lab scientists, and corporate R&D teams—while the communities, workers, policymakers, and future generations most affected by the resulting materials are systematically excluded [13, 14]. Objective functions are not politically neutral algorithms; they are instruments of power that allocate research effort, funding, and ultimately material resources. When a small group of specialists chooses to minimize formation energy or maximize ionic conductivity without broader consultation, they are effectively deciding whose values will shape the built environment for decades to come.

Current practice treats objective selection as an internal technical matter. Butler and co-workers review machine-learning applications in materials without once mentioning stakeholder input [4]. At the same time, Gómez-Bombarelli *et al.* [7] present their generative framework as a purely data-driven exercise in molecular design. The absence of participatory mechanisms is not an oversight but a structural feature: the optimization pipeline offers no natural entry point for non-expert voices. A community living downstream from a proposed cobalt-free cathode material has no formal channel to insist that supply-chain ethics be encoded as a hard constraint; a policymaker concerned with critical-mineral dependence has no seat at the table when the loss function is being written. The result is a technocratic monopoly on value definition that mirrors broader patterns of expert closure documented in science-and-technology-studies literature [10, 23].

This deficit becomes especially acute when materials cross into societal deployment. A solid-state electrolyte optimized for maximum conductivity may rely on lithium extraction practices that affect Indigenous territories or exacerbate water scarcity in arid regions. Yet, the objective function contains no term that registers those impacts because the affected parties were never consulted during its formulation [28]. Similarly, thermoelectric materials pushed toward higher zT through rare-earth doping may solve waste-heat recovery in data centers while creating new waste streams that burden recycling infrastructures in the Global South. The decision about which trade-offs are acceptable was made upstream, in the privacy of the code notebook, by researchers whose lived experience and incentive structures rarely overlap with those of the communities downstream.

Russell's call for human-compatible artificial intelligence emphasizes that value alignment requires explicit negotiation with the humans whose values are at stake [2]. Materials AI has ignored this imperative, treating alignment as an internal consistency problem rather than a sociopolitical one. The critique, therefore, demands a shift from implicit to explicit governance: objective functions must emerge from participatory processes that include diverse stakeholders at the earliest stages of design. Such processes might involve deliberative workshops, citizen assemblies, or structured value-elicitation protocols that translate lived experiences into mathematical constraints. Only then can the field claim legitimacy in deciding what materials the future deserves [18].

The democratic deficit is not merely procedural; it is epistemic. Diverse stakeholders bring forms of knowledge—local ecological expertise, occupational health insights, intergenerational justice perspectives—that are invisible to purely technical optimization. By excluding those knowledges, materials AI narrows its own epistemic base and produces brittle

solutions that fail when confronted with the full complexity of real-world contexts [11, 21]. The present analysis, therefore, insists that “who decides” is not a peripheral question of science policy but a core determinant of the normative content of objective functions themselves. Until democratic input is institutionalized, materials AI will remain an elite project whose value commitments reflect the priorities of the few rather than the needs of the many [3].

Consequences for Materials Discovery

The unexamined normative assumptions embedded in materials AI objective functions generate four interlocking consequences that undermine the very promise of accelerated discovery.

Narrow discovery

Because optimization is channeled through a limited set of value-laden targets, the search space collapses around materials that satisfy those targets while discarding vast regions of potentially valuable chemical space [4, 5]. Entire families of metastable or topologically protected materials are never generated simply because they score poorly on the chosen scalar. The pipeline, therefore, produces a narrower, more homogeneous portfolio of candidates than physics itself would allow.

Value conflicts unresolved

Trade-offs between competing societal priorities—efficiency versus sustainability, performance versus safety—are either hidden inside weighted sums or ignored entirely [2]. When conflicts surface downstream, they appear as unexpected engineering failures rather than as the predictable outcome of upstream value choices. The field is left to retrofit ethical constraints after the fact, an approach that is both inefficient and ethically questionable.

Democratic deficit

As detailed above, the exclusion of affected stakeholders from objective formulation means that materials discovery proceeds without the consent or input of those who will live with the consequences [13, 14]. Public trust erodes when technologically impressive materials later reveal hidden social or environmental costs that could have been anticipated through inclusive design.

Path-dependent values

Early choices of objective functions become self-reinforcing: successful papers cite the same targets, funding calls reward them, and subsequent models are trained on data distributions shaped by them [7]. A generation of researchers, therefore, inherits a value monoculture that is increasingly difficult to escape, locking the field onto trajectories that may prove suboptimal or even harmful once broader societal priorities shift.

Collectively, these consequences transform materials AI from a liberatory technology into a mechanism that reproduces and amplifies the very limitations it claims to overcome [3, 20].

Alternative Approaches

Five constructive alternatives can move materials AI toward explicit, reflexive, and pluralistic objective design.

Explicit value statements

Every publication and every code repository should be required to include a concise “value declaration” that articulates the normative commitments encoded in the objective function, the trade-offs accepted, and the stakeholders implicitly prioritized [3, 24]. Such declarations would transform hidden normativity into transparent accountability.

Multi-objective optimization with value articulation

Rather than scalarize competing objectives through arbitrary weights, pipelines should maintain Pareto fronts that explicitly surface trade-offs among incommensurable values, allowing downstream decision-makers to choose according to context-specific priorities.

Participatory objective design

Structured deliberative processes—citizen panels, stakeholder workshops, or hybrid human-AI value-elicitation protocols—should be convened before objective functions are finalized, ensuring that affected communities help shape the very loss landscapes the algorithms will traverse [18].

Value-sensitive optimization

Objective functions can be architected with modular "value modules" that can be activated or deactivated according to context, each module carrying its own explicit normative rationale and uncertainty bounds [2]. This modularity would allow the same underlying model to serve sustainability-focused or performance-focused campaigns without requiring a complete redesign.

Reflexive objective auditing

Periodic third-party audits—modeled on algorithmic impact assessments—should evaluate whether the deployed objectives continue to align with evolving societal values, triggering revisions when path dependence or new ethical horizons render earlier choices obsolete [23].

Together, these alternatives reframe objective-function design as an ongoing sociotechnical practice rather than a one-time technical specification [1, 3].

Table 2 translates the manuscript's critique into a corrective framework by linking each diagnosed problem to its structural source, its persistence mechanism, and a corresponding governance response.

Table 2. From hidden normativity to reflexive governance: a corrective framework for objective-function design in materials AI

Problem diagnosed in the manuscript	Structural source in current practice	Why the problem persists	Corrective design principle	Concrete governance mechanism	Expected improvement
Hidden normativity	Objective functions are framed as purely technical parameters	Mathematical formalism obscures upstream value choices	Make normative commitments explicit at the point of objective definition	Mandatory value declaration in papers, repositories, and supplementary methods	Greater transparency and contestability of optimization choices
Value monoculture	Community incentives reward repeated optimization of the same	Publication, benchmarking, and dataset cultures reinforce a narrow value horizon	Build value pluralism into optimization design	Pareto-front reporting across scientific, environmental, and societal dimensions	Broader discovery portfolios and reduced overconcentration on canonical metrics

	benchmark properties				
Measurability trap	Only easily computable properties become optimization targets	Data availability and tractable simulation bias what enters the loss function	Distinguish measurable proxies from ultimate societal or scientific aims	Multi-fidelity objective hierarchies, proxy warnings, uncertainty annotations, deferred constraints	Better alignment between what is optimized and what actually matters in deployment
Democratic deficit	Objective selection is controlled by technical elites	Stakeholders lack formal entry points into upstream design decisions	Treat objective formulation as a participatory sociotechnical process	Stakeholder workshops, deliberative panels, and structured value elicitation before model deployment	Increased legitimacy, broader epistemic input, and stronger alignment with affected communities
Path-dependent values	Early objective choices shape datasets, benchmarks, and future funding priorities	Success reproduces itself through training data, citations, and infrastructure	Subject objectives to periodic re-evaluation rather than treating them as fixed	Third-party objective audits and scheduled redesign reviews	Reduced lock-in and greater adaptability to changing societal priorities

Conclusion

This critical critique has demonstrated that objective functions in materials AI are never neutral technical choices; they are vehicles of normative judgment that encode contested visions of what constitutes a good material. From the hidden normativity concealed behind scalar loss functions, through the value monoculture that narrows discovery horizons, the measurability trap that substitutes computable proxies for genuine relevance, to the democratic deficit that excludes those most affected, the field has operated under the illusion of value-free optimization. The consequences—narrow trajectories, unresolved conflicts, eroded legitimacy, and path-dependent futures—now threaten to undermine the societal promise of accelerated materials innovation.

The alternative approaches outlined above offer a pathway forward: explicit value articulation, pluralistic optimization, participatory governance, modular sensitivity, and reflexive auditing can transform materials AI into a genuinely reflexive practice. The field must therefore abandon the comforting fiction that objective functions are merely mathematical conveniences and instead treat their design as a site of ethical and political deliberation. Only by making the normative content of optimization visible and contestable can materials science fulfill its responsibility to deliver not merely faster discovery but discovery that is aligned with the full spectrum of human and planetary needs. The collapse of the fact/value dichotomy is not a philosophical curiosity; it is the central design challenge facing materials AI today.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 26 Dec 2024 Revised: 02 Feb 2025 Accepted: 14 Mar 2025
Published online: 18 July 2025

References

Chakraborty S. The fact/value dichotomy: Revisiting putnam and habermas. Philosophia. 2019;47(2):369-86.

Russell S. Human-compatible artificial intelligence. Hum Like Mach Intell. 2022;1:3-22.

Mühlhoff R, Ruschemeier H. Updating purpose limitation for AI: A normative approach from law and philosophy. Int J Law Inf Technol. 2025;33:eaaf003.

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.

Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, et al. Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci. 2018;4(2):268-76.

Cercone N, McCalla G. Artificial intelligence: Underlying assumptions and basic objectives. J Am Soc Inf Sci. 1984;35(5):280-90.

Kaufman BE. The social welfare objectives and ethical principles of. In: The ethics of human resources and industrial relations. Champaign: LERA; 2005. p. 23-59.

Giannini F, Diligenti M, Maggini M, Gori M, Marra G. T-norms driven loss functions for machine learning. Appl Intell. 2023;53(15):18775-89.

De Gregorio G. The normative power of artificial intelligence. Ind J Global Legal Stud. 2023;30:55.

Trehern W, Ortiz-Ayala R, Atli KC, Arroyave R, Karaman I. Data-driven shape memory alloy discovery using artificial intelligence materials selection (AIMS) framework. Acta Mater. 2022;228:117751.

Madika B, Saha A, Kang C, Buyantogtokh B, Agar J, Wolverton C, et al. Artificial intelligence for materials discovery, development, and optimization. ACS Nano. 2025;19(30):27116-58.

Sadrossadat E, Basarir H, Karrech A, Elchalakani M. Multi-objective mixture design and optimisation of steel fiber reinforced UHPC using machine learning algorithms and metaheuristics. Eng Comput. 2022;38(Suppl 3):2569-82.

Prasittisopin L. Artificial intelligence-driven materiality: A systematic review and perspective. Eng Sci. 2025;35:1587.

Kumar S, Choudhury S. Normative ethics, human rights, and artificial intelligence. AI Ethics. 2023;3(2):441-50.

Wang J, Wang Y, Chen Y. Inverse design of materials by machine learning. Materials. 2022;15(5):1811.

Gabriel I. Artificial intelligence, values, and alignment. Minds Mach. 2020;30(3):411-37.

Reiss J. Fact-value entanglement in positive economics. J Econ Methodol. 2017;24(2):134-49.

Galos JL, Kivy MB, Grunenfelder L, Nomura KI, Saal JE. A critical review of approaches to teaching artificial intelligence in undergraduate materials engineering. In: 2025 ASEE Annual Conference Exposition; 2025.

Çullhaj F. Complications (complexity) between normative and descriptive: A challenge for clarity. Balk J Philos. 2022;14(1):65-72.

Li JP, Polovina N, Konur S. A review of AI-driven engineering modelling and optimization: Methodologies, applications and future directions. Algorithms. 2026;19(2).

Raap M, Preuß M, Meyer-Nieberg S. Moving target search optimization–a literature review. Comput Oper Res. 2019;105:132-40.

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.

Bagheri M, Bagheritabar M, Alizadeh S, Parizi MS, Matoufinia P, Luo Y. Machine-learning-powered information systems: A systematic literature review for developing multi-objective healthcare management. Appl Sci. 2024;15(1):296.

Niiniluoto I. Assessing value-laden technology. Rev Guillermo Ockham. 2024;22(2):5-17.

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.

Johnson JP. The fact-value question in early modern value theory. In: Value theory in philosophy and social science (RLE Social Theory). Routledge; 2014. p. 64-71.

Billard TJ. Theory and/as normative assumptions in political communication research. Political Communication Report. 30. Available from: https://nbn-resolving.org/urn:nbn:de:0168-ssoar-98561-6