ORIGINAL RESEARCH **Open access**

Algorithmic Attention as Scientific Bias: A Conceptual Analysis for Materials AI

Lucas Meyer[1*], Stefan Braun[1], Anna Schmid[2]

Abstract

The rapid integration of machine learning and artificial intelligence into materials science has introduced powerful capabilities for predicting, screening, and discovering new materials. Yet this integration also engenders a distinctive form of bias that operates not merely through skewed training data but through the mechanisms by which models allocate and distribute attention across chemical, structural, and property spaces. This paper conceptualizes "algorithmic attention" as a form of scientific bias that manifests in materials AI systems, shaping which phenomena receive emphasis, which regions of materials space are explored, and ultimately which knowledge claims gain epistemic legitimacy within the field. Attention is interpreted here as the patterned prioritization embedded in model architectures, loss functions, data sampling strategies, and iterative feedback loops between prediction and experiment. The analysis explores how such attention dynamics amplify existing data imbalances, create self-reinforcing discovery loops, misalign interpretive authority between model outputs and domain expertise, complicate validation of uncertain predictions, steer research trajectories through hidden optimization priorities, and pose system-level challenges for epistemic reliability and governance. Drawing on recent literature in materials informatics, bias in machine learning, and philosophy of data-driven science, the paper develops an integrative conceptual framework that treats algorithmic attention as an emergent property of socio-technical knowledge systems rather than a purely technical artifact. This framing highlights trade-offs between predictive scalability and epistemic pluralism, underscoring the need for reflective practices that render attention mechanisms more visible and contestable within materials discovery workflows.

Keywords Materials AI, Algorithmic attention, Scientific bias, Epistemic steering, Feedback loops, Interpretability

*Correspondence:
Lucas Meyer
lucas.meyer@gmail.com
[1] Department of Materials Modeling and Artificial Intelligence, Faculty of Engineering, ETH Zurich, Zurich, Switzerland
[2] Department of Data-Driven Materials Science, Faculty of Engineering, University of Bern, Bern, Switzerland

Introduction

Materials science stands at a pivotal juncture in its history. The convergence of high-throughput computation, large-scale experimental databases, and advanced machine learning has enabled modes of inquiry that were previously unimaginable: rapid screening of vast chemical spaces, inverse design of target properties, and autonomous iteration between prediction and measurement [1-4]. These developments promise accelerated discovery of functional materials critical to energy storage, catalysis, quantum technologies, and sustainable manufacturing [5-7]. Yet the same technological infrastructure that enables such acceleration simultaneously introduces novel forms of distortion into the scientific process.

Conventional discussions of bias in materials AI have largely centered on imbalances in training datasets—over-representation of certain crystal structures, elemental compositions, or property regimes that reflect historical research priorities rather than intrinsic physical abundance [8-10]. While data skew remains an important concern, it captures only

part of the phenomenon. Increasingly, bias arises not solely from what data exist, but from how models learn to attend to those data: which features, interactions, correlations, or subspaces receive disproportionate weighting during training, inference, and active learning cycles [11, 12]. This patterned allocation of computational focus—what this analysis terms *algorithmic attention*—functions as a subtle yet pervasive form of scientific bias. It influences not only what materials are predicted to be promising, but which questions are posed, which explanations are privileged, and which trajectories of discovery are reinforced over others.

The concept of attention here draws from its technical usage in transformer architectures and graph neural networks, where attention mechanisms dynamically assign importance weights to different parts of an input representation [1]. Yet it extends beyond model internals to encompass broader socio-technical dynamics: the interplay among model design choices, the curation of training corpora, the selection of objective functions, human interpretation of outputs, and the feedback loops that connect AI-generated hypotheses to subsequent experiments [12, 13-15]. When models repeatedly attend to certain regions of phase space—say, oxide perovskites over chalcogenides, or high-symmetry structures over disordered systems—these patterns become inscribed in the field's collective knowledge base, subtly reshaping scientific understanding [16, 17].

Such attention-driven bias differs from classical statistical bias in several respects. It is often emergent rather than intentional, distributed across multiple layers of a modeling pipeline, and resistant to correction through dataset balancing alone [11, 18]. Moreover, because materials discovery operates in high-dimensional, sparsely sampled spaces, small differences in attention allocation can compound exponentially, leading to pronounced divergence between model-guided exploration and what might have been pursued under alternative epistemic priorities [7, 19]. The consequence is not merely an inaccurate prediction but a narrowing of the discipline's conceptual and empirical horizons.

This conceptual analysis seeks to articulate algorithmic attention as a coherent form of scientific bias specific to data-intensive materials research. It proceeds in three principal movements. First, it synthesizes relevant strands from the recent literature on materials informatics, machine learning fairness, and epistemology of computational science to identify recurring themes of amplification, feedback, interpretation, validation, steering, reliability, integration, and governance. Second, it proposes an integrative framework that positions algorithmic attention at the center of a dynamic system linking model behavior, data practices, human judgment, and institutional incentives. Finally, it offers interpretive reflections on the epistemic and ethical stakes of attending differently to materials AI.

The analysis remains deliberately conceptual, refraining from empirical claims, new propositions, or prescriptive remedies. Instead, it aims to reframe existing observations within a unified interpretive lens that renders visible the subtle ways in which attention mechanisms mediate the production of scientific knowledge in an increasingly AI-mediated field. By doing so, it seeks to contribute to a more reflexive practice of materials discovery, one that acknowledges the inseparability of technical design and epistemic consequence.

Theoretical Background and Literature Synthesis

Amplification through Attention Mechanisms: Contemporary machine learning systems frequently rely on architectures—such as graph neural networks, transformers, and message-passing networks—that explicitly incorporate attention layers [1, 2, 4]. These mechanisms allow models to focus selectively on chemically or structurally salient features while downweighting others. Although designed to improve predictive accuracy, such selective focus inherently amplifies pre-existing asymmetries in the data landscape. Regions of materials space that are densely sampled in historical databases receive sustained attention, while sparsely populated or historically neglected regions are systematically de-emphasized [8, 9]. Recent work has illustrated how scaling laws in deep learning exacerbate this dynamic: as models grow larger and are trained on ever-larger datasets, small initial imbalances in attention distribution become dramatically magnified [1, 16]. The result is a form of bias that operates downstream of data collection, rooted in the inductive preferences encoded within model architectures themselves.

Feedback Loops between Prediction and Experiment: A defining characteristic of modern materials discovery workflows is the iterative coupling of model prediction and experimental validation, often realized through active learning or autonomous laboratories [7, 13, 19]. In these loops, model outputs influence which experiments are performed next, and experimental outcomes, in turn, update the training distribution. Algorithmic attention plays a critical mediating role here. When attention is disproportionately allocated to high-confidence or high-utility regions (as defined by acquisition functions), the feedback loop tends to reinforce exploration of already well-sampled subspaces, creating path-dependent discovery trajectories [12, 15]. The literature on self-reinforcing biases in machine learning highlights analogous dynamics in other domains, where predictive models progressively entrench their inductive priors [11, 18, 20]. In materials science, this manifests as a narrowing of the explored phase space over successive iterations, even when the initial dataset was intended to be broadly representative.

Interpretation and Misalignment of Authority: Model outputs in materials AI are rarely self-interpreting. Human experts must translate numerical predictions, attention maps, or feature importance scores into chemical insight or physical mechanism [3, 9]. Yet the attention patterns learned by models often diverge from domain-grounded notions of scientific relevance. An attention head may assign high weight to geometric descriptors that correlate spuriously with a target property within the training manifold but lack causal significance outside it [17]. When such patterns are interpreted as evidence of underlying scientific truth, a form of epistemic misalignment arises: algorithmic salience is mistaken for explanatory salience [12]. Recent reflections on explainable AI in materials design underscore this tension, noting that post-hoc interpretability tools may themselves introduce secondary distortions by privileging computationally accessible explanations over mechanistically valid ones [9].

Validation Challenges in Uncertain Regimes: Materials discovery frequently targets regimes far from existing data—extrapolative rather than interpolative prediction. In such regimes, model uncertainty is high, yet attention mechanisms often remain confident, focusing on familiar patterns even when they are inappropriate [7]. Validation becomes particularly challenging because ground truth is absent or prohibitively expensive to obtain. The literature reveals a recurring tension between reliability metrics computed within the training distribution and actual performance in out-of-distribution discovery tasks [8, 16]. Algorithmic attention contributes to this difficulty by masking uncertainty: models may attend strongly to a small subset of features that yield low training error but fail to generalize, leading to overconfident extrapolations that resist falsification [11].

Steering through Hidden Priorities: Every modeling choice embeds priorities—whether explicit (e.g., multi-objective loss terms) or implicit (e.g., architectural inductive biases, regularization strategies). These priorities steer discovery toward certain classes of materials or properties at the expense of others [12, 21]. In materials AI, steering often occurs invisibly: attention heads learn to prioritize symmetry-adapted features, local bonding motifs, or electronic descriptors that reflect the optimization landscape rather than intrinsic scientific importance [2, 4]. Governance-oriented discussions in adjacent fields emphasize that such hidden steering constitutes a form of power that shapes research agendas without explicit deliberation [2, 20, 22].

Reliability versus Scalability Trade-offs: Larger models and broader datasets promise greater predictive coverage, yet they also intensify attention concentration on dominant patterns within the data [1, 16]. Smaller, more interpretable models may preserve epistemic pluralism but sacrifice scalability [3]. This trade-off is not merely technical but epistemic: reliability in uncertain, high-stakes discovery contexts may require deliberate diversification of attention, even at the cost of peak predictive performance [12, 15].

System-Level Integration and Governance: Algorithmic attention cannot be isolated from the broader ecosystem of data infrastructures, publication incentives, funding priorities, and peer-review practices [13-15]. Governance frameworks must therefore address attention at the system level—through transparent documentation of attention patterns, participatory design of objective functions, or institutional mechanisms that incentivize exploration of under-attended regions [21, 23]. Emerging discussions on responsible AI in science stress the need for epistemic accountability structures capable of rendering attention dynamics visible and contestable [12, 24]. Taken together, these strands indicate

that algorithmic attention operates as a distributed bias spanning model design, workflow dynamics, interpretation, and institutional context **(Table 1)**.

Table 1. Forms of algorithmic attention bias across the materials AI pipeline

Pipeline layer	Mechanism of attention allocation	Resulting bias pattern	Epistemic consequence
Model architecture	Attention heads prioritize recurring geometric, chemical, or electronic motifs	Over-weighting of historically dominant structure–property correlations	Reinforcement of inductive priors mistaken for physical salience
Loss functions and optimization	Objective functions reward performance in dense regions of training space	Attention concentrates on high-confidence regimes	Narrowing of explored chemical and structural diversity
Data–model feedback loops	Active learning and experiment selection reinforce prior attention distributions	Path-dependent exploration trajectories	Progressive contraction of the discovery horizon
Interpretability interfaces	Saliency maps and attention scores frame explanations	Algorithmic relevance conflated with scientific explanation	Misalignment between statistical importance and causal meaning
Validation practices	Benchmarks emphasize in-distribution accuracy	Attention remains confident under extrapolation	Masked uncertainty and overconfidence in frontier regimes
Institutional incentives	Funding, publication, and benchmarking norms reward fast convergence	System-level steering of attention	Marginalization of low-attention but mechanistically plausible domains

Proposed conceptual framework

This analysis proposes a conceptual framework that positions *algorithmic attention* as the central mediating process through which bias emerges and propagates in materials AI systems. Rather than treating attention as a purely technical component, the framework interprets it as an epistemic operator that links four interdependent layers: (1) model-internal attention dynamics, (2) data-model feedback loops, (3) human-AI interpretive interfaces, and (4) institutional steering structures.

At the core, model-internal attention operates as a distributional mechanism: it reallocates representational resources across chemical, structural, and property dimensions according to learned weights. These weights reflect not only statistical regularities but also the cumulative effect of architectural choices, loss landscapes, and training curricula. The framework views this core as inherently selective and therefore value-laden—attention is never a neutral allocation but always a form of prioritization with epistemic consequences.

Surrounding this core lie bidirectional feedback loops. Predictions weighted by attention influence experimental selection (via active learning or human triage), while new measurements reshape the training manifold, further tuning attention patterns. The framework highlights the self-amplifying nature of these loops: attention begets data, which in turn begets attention, creating trajectories that are path-dependent and resistant to external correction.

Encircling the feedback layer is the interpretive interface, where domain experts encounter model outputs. Here, attention maps, saliency scores, and uncertainty estimates are translated into scientific narratives. The framework emphasizes

misalignment risks: algorithmic attention may highlight correlations that are statistically robust yet scientifically incidental, while obscuring causally relevant but low-attention features. This layer introduces a second-order bias—bias about which model behaviors count as meaningful.

Finally, the institutional layer encompasses the broader socio-technical environment: publication norms that reward high-performance benchmarks, funding structures that favor rapid discovery, and community standards for reproducibility and interpretability. These forces exert long-range steering on attention by shaping which modeling paradigms are rewarded and which are marginalized.

The framework can be visualized as a nested, dynamic system in **Figure 1**.

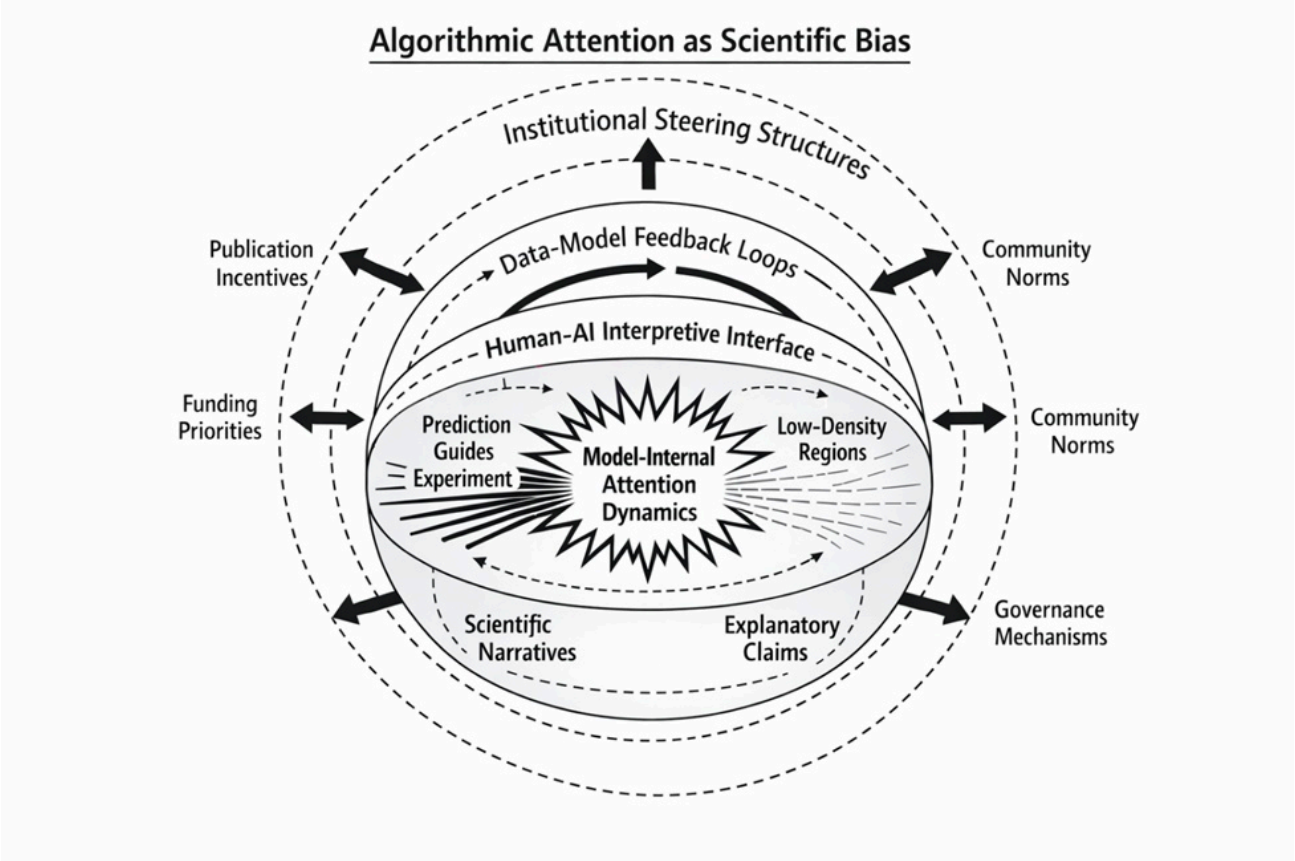

Figure 1. Systems-level model of algorithmic attention bias in materials AI

Figure 1 illustrates the emergent, system-level nature of algorithmic attention as a source of scientific bias in materials AI. A nested circular diagram shows how model-internal attention, preferentially directed toward high-density data regions, is locked in via Data-Model Feedback Loops that steer experiments and update training. This technical bias is filtered through the Human-AI Interpretive Interface, where selective attention patterns are translated into scientific narratives. The entire system is framed and reinforced by Institutional Steering Structures (e.g., funding priorities, publication incentives). Bidirectional dashed lines across all layers demonstrate the co-constitutive feedback that entrenches attention-driven bias as a systemic epistemic issue, rather than a mere technical flaw.

This framework does not prescribe specific interventions but offers an integrative lens for interpreting how attention operates across scales—from individual attention heads to field-level discovery patterns. It invites reflection on the epistemic costs of attending to data-driven materials science strongly and narrowly versus weakly and broadly.

Analytical implications

The framework advanced here invites a series of interpretive shifts in how bias is understood within materials AI. Rather than confining bias to input data distributions or output disparities, it relocates the locus of distortion to the ongoing, dynamic process of attention allocation. This relocation carries several analytical consequences.

First, it reframes amplification as an intrinsic property of learning rather than an accidental artifact. Attention mechanisms do not merely reflect data skew; they actively reshape the effective manifold that the model "sees" over training epochs and inference steps. As models scale in parameter count and dataset size, attention tends to concentrate ever more sharply on statistically dominant patterns, producing a form of epistemic gravitational pull toward familiar chemical families, structural prototypes, and property regimes [1, 7, 16]. The implication is that discovery acceleration is coupled with a systematic contraction of conceptual breadth: the faster and larger the system, the narrower the attended horizon becomes, unless deliberate countermeasures are embedded at the architectural or training level. The principal analytical consequences of treating attention as a form of scientific bias are summarized in **Table 2**, which maps specific attention dynamics to their epistemic effects and broader scientific stakes.

Table 2. Analytical implications of algorithmic attention as scientific bias in materials AI

Attention dynamic	Analytical shift introduced	Epistemic risk identified	Implication for materials discovery
Amplification via attention concentration	Bias reframed as intrinsic to learning dynamics	Conceptual contraction despite predictive scaling	Accelerated discovery coupled with reduced epistemic diversity
Feedback-driven path dependence	Optimization cycles are seen as epistemic steering mechanisms	Early design choices lock in long-term trajectories	Irreversibility of discovery agendas
Interpretive misalignment	Attention treated as rhetorical authority	Algorithmic salience mistaken for explanation	Canonization of spurious mechanisms
Extrapolative attention persistence	Validation reframed as reflexive rather than metric-driven	Underestimated risk in frontier predictions	False confidence in novel materials regimes
Hidden optimization priorities	Technical design exposed as value-laden	Undeliberated agenda-setting	Implicit governance through model architecture
System-level reinforcement	Bias is understood as socio-technical	Local debiasing rendered ineffective	Need for coordinated, multi-layer accountability

Second, feedback loops emerge not as neutral optimization cycles but as carriers of path dependence. Each iteration in an active-learning pipeline reinforces the attention distribution inherited from the previous round, entrenching initial inductive biases into the evolving knowledge base [7, 12, 19]. This dynamic suggests that early modeling choices—loss weighting, acquisition strategy, initial architecture—exert disproportionate long-term influence on which regions of materials space are deemed worthy of exploration. The framework thus highlights a temporal asymmetry: attention patterns established in the first few dozen cycles can lock in trajectories that persist across years of research, even as new data accumulate.

Third, the interpretive interface layer reveals a subtle inversion of epistemic authority. When attention maps or saliency analyses are presented alongside predictions, they often function less as neutral diagnostic tools and more as rhetorical devices that lend scientific legitimacy to model decisions [3, 9]. The risk is not only misinterpretation but a gradual transfer

of explanatory prerogative from domain-grounded reasoning to computationally salient patterns. Over time, concepts that survive repeated high-attention reinforcement may acquire the status of canonical mechanisms, while alternatives that occupy low-attention subspaces remain marginal or invisible, even when mechanistically plausible [12, 17].

Fourth, validation in extrapolative regimes takes on a distinctly reflexive character. Because attention often remains concentrated in high-confidence (i.e., high-density) regions even when uncertainty is high, standard uncertainty quantification metrics may systematically understate risk in discovery-oriented tasks [8]. The framework suggests that reliability cannot be assessed solely through cross-validation or calibration scores; it requires scrutiny of how attention behaves under distributional shift. A model that attends robustly across sparse regimes may appear less accurate on benchmark tasks yet prove more epistemically trustworthy for genuine frontier exploration.

Fifth, steering through hidden priorities exposes a deeper entanglement between technical design and research agenda-setting. Optimization choices that appear purely instrumental—regularization strength, multi-task weighting, attention-head count—implicitly encode value judgments about which materials classes, which performance metrics, and which discovery speeds merit priority [12, 15, 21]. The analytical implication is that materials AI is never value-free; every deployed system enacts a particular vision of what counts as scientific progress, often without explicit articulation or contestation.

Finally, the nested, system-level character of the framework underscores that no single layer can be reformed in isolation. Attempts to debias attention solely through dataset augmentation or post-hoc reweighting are likely to be swamped by feedback reinforcement and institutional steering [11, 18, 24]. Meaningful change, therefore, demands coordinated intervention across model architecture, workflow design, interpretive practices, and governance structures. This insight aligns with emerging calls for holistic responsibility frameworks in data-driven science [12, 23].

Results and Discussion

Taken together, these analytical implications reveal algorithmic attention as more than a technical challenge; it constitutes a distinctive mode of scientific bias, uniquely adapted to the high-dimensional, sparse, and iterative nature of materials discovery. Unlike classical selection bias or confounding, which can often be mitigated through statistical adjustment, attention-mediated bias is distributed, recursive, and partially opaque even to system designers. It operates at the intersection of computation and cognition, quietly shaping not only what is discovered but how discovery itself is conceptualized within the community.

This perspective complicates prevailing narratives of progress in materials informatics. While scaling has delivered remarkable predictive capabilities [1, 16], it simultaneously intensifies attention concentration, raising questions about whether acceleration is purchased at the expense of epistemic diversity. The framework does not deny the utility of powerful models; rather, it suggests that their benefits are unevenly distributed across the conceptual landscape of materials science. Regions that align with prevailing attention patterns benefit from rapid iteration and confirmation, while those that fall outside receive delayed or indirect attention, if any at all.

The discussion also surfaces inherent trade-offs. Greater predictive power through attention concentration enhances efficiency in well-mapped domains but risks premature convergence in frontier areas. Deliberate diversification of attention—through ensemble methods, curiosity-driven acquisition, or architecturally enforced sparsity—may preserve pluralism at the cost of peak performance on standard benchmarks [3, 19]. Similarly, rendering attention more transparent through advanced interpretability tools may improve accountability yet introduce new interpretive distortions if those tools themselves reflect biased design choices [9].

Institutionally, the framework points toward the need for mechanisms that make attention dynamics legible and contestable. Current publication and funding cultures often reward high hit rates in narrow property spaces, indirectly incentivizing attention concentration [13, 15]. Shifting evaluative norms toward measures of exploratory breadth,

uncertainty-aware reporting, or attention-diversity metrics could help counterbalance these pressures, though such shifts would require sustained community coordination.

Ethically, the opacity and path-dependence of attention allocation raise questions of responsibility. Who bears accountability when a model attends overwhelmingly to one class of materials, steering entire subfields away from alternatives? The distributed nature of the system—spanning model developers, database curators, experimentalists, reviewers, and funders—diffuses responsibility. Yet, the epistemic consequences remain concentrated in foreclosed possibilities and entrenched assumptions [2, 22, 24].

Ultimately, recognizing algorithmic attention as scientific bias invites a more reflexive posture toward AI-mediated discovery. It calls for ongoing critical engagement with the question of what, and how, materials science attends when guided by computational systems.

Conclusion

This conceptual analysis has sought to illuminate algorithmic attention as a coherent and consequential form of bias within materials AI. By interpreting attention not as a neutral computational primitive but as an epistemic operator embedded in layered socio-technical dynamics, the framework reveals how seemingly technical choices reverberate through prediction, experimentation, interpretation, validation, and agenda-setting.

The central insight is that bias in data-driven materials science is increasingly produced through the patterned ways models learn to see and prioritize. These patterns, once established, propagate through feedback loops, solidify via human interpretation, and are reinforced by institutional structures, creating self-sustaining cycles of discovery that favor certain regions of knowledge over others. While the acceleration enabled by modern AI is undeniable, the framework cautions that such acceleration is never epistemically neutral; it enacts particular ways of attending that carry long-term consequences for the breadth, depth, and pluralism of materials understanding.

By centering attention as the mediating process, the analysis offers a unifying lens for disparate observations in the literature—from the scaling-induced magnification of data skew to active-learning path dependence, to interpretability misalignment and governance challenges. It does not propose ready solutions but establishes a conceptual vocabulary that renders these dynamics more visible and therefore more amenable to deliberate reflection and adjustment.

As materials AI continues to mature, attending to attention itself may prove essential to preserving the creative and critical openness that has long characterized scientific inquiry. Future work could extend this framing by examining specific attention patterns across model families, tracing their evolution through discovery campaigns, or exploring institutional designs that foster more pluralistic modes of algorithmic attention. For now, recognizing that how models attend is inseparable from what science comes to know remains a necessary step toward a more responsible integration of artificial intelligence into materials discovery.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 20 Jun 2021 Revised: 20 Aug 2021 Accepted: 14 Sep 2021
Published online: 18 January 2022

References

Lu T, Cibralic B, Lark SL, Tucker AG, Albers DJ, Seltzer M. Moving beyond “algorithmic bias is a data problem”. Patterns. 2021;2(4):100240.
https://doi.org/10.1016/j.patter.2021.100240.

Chávez-Angel E, Mugarza A, Sotomayor Torres CM. Applied artificial intelligence in materials science and material design. Adv Intell Syst. 2025;7(7):2400986.
https://doi.org/10.1002/aisy.202400986.

Varshney V, Boland JJ, Roy AK, Wheeler RL, Brown KA. Explainable artificial intelligence (XAI) for material design and engineering applications: a quantitative computational framework. Mater Sci Des. 2025;1(1):1-12.
https://doi.org/10.1002/msd2.70017.

Gyftakis KN, Charitidis P. Advancing materials discovery through artificial intelligence. Appl Mater Today. 2025;41:101935.
https://doi.org/10.1016/j.apmt.2025.101935.

Pamla B, Molokwane T, Mpuru TJ, Segalo T, Kyesmen PI, Diale M. Application of artificial intelligence in materials science, with a special focus on fuel cells and electrolyzers. J Energy Storage Mater. 2024;8:100318.
https://doi.org/10.1016/j.ensm.2024.100318.

Sun H, Peng S, Song K, Song M, Yu PS. Bias of AI-generated content: an examination of news produced by large language models. Sci Rep. 2024;14(1):4887.
https://doi.org/10.1038/s41598-024-55686-2.

Spence N. A model for shared epistemic agency. In: Proceedings of the 16th International Conference of the Learning Sciences (ICLS 2022). Int Soc Learn Sci; 2022. p. 799-806.

Ogunleye O, Atarah SA, Fischer CC, Rinderspacher BC, Chung PW. Materials informatics: a review of AI and machine learning tools, platforms, data repositories, and applications to architectured porous materials. Mater Des. 2025;242:113057.
https://doi.org/10.1016/j.matdes.2025.113057.

Fischer CC, Rinderspacher BC, Chung PW. Effects of data bias on machine-learning–based material discovery using experimental property data. Mater Discov. 2022;20:101234.
https://doi.org/10.1016/j.md.2022.101234.

Himanen L, Geurts A, Foster AS, Rinke P. Data-driven materials science: status, challenges, and perspectives. Adv Sci. 2019;6(21):1900808.
https://doi.org/10.1002/advs.201900808.

Ogunleye O, Chung PW. Machine learning-driven material intelligence research and development. Nanoscale Res Lett. 2025;20(1):95.
https://doi.org/10.1186/s11671-025-04009-5.

Li Y, Chen C, Zuo Y, Ye W, Ong SP. A decision architecture for epistemic prioritization: machine learning at the intersection of technology and society. J Mater Inform. 2025;5(1):13.
https://doi.org/10.20517/jmi.2025.13.

Chen Z, He D, Zhao X, Zhao S, Zhang J, Pan M, et al. Unleashing the power of AI in science: key considerations for materials data preparation. Sci Data. 2024;11(1):945.
https://doi.org/10.1038/s41597-024-03821-z.

Polak MP, Morgan D. 1.5 million materials narratives generated by chatbots. Sci Data. 2024;11(1):1019.
https://doi.org/10.1038/s41597-024-03886-w.

Li N, Tian X, Cui Y, Wei C, Song X, Zhang C, et al. A review of the use of AI in the mining industry: insights and ethical considerations for multi-objective optimization. Resour Policy. 2024;89:104646.
https://doi.org/10.1016/j.resourpol.2024.104646.

Merchant A, Batzner S, Schoenholz SS, Aykol M, Cheon G, Cubuk ED. Scaling deep learning for materials discovery. Nature. 2023;624(7990):80-5.
https://doi.org/10.1038/s41586-023-06735-9.

Nematov D, Hojamberdiev M. AI-driven materials design: a mini-review. arXiv. 2025:arXiv:2502.02905.
https://doi.org/10.48550/arXiv.2502.02905.

Ning Y, Ghosh M, Quer G, Nilsen W, Yadav A, Vinterbo SA, et al. A survey of recent methods for addressing AI fairness and bias in biomedicine. J Biomed Inform. 2024;154:104646.
https://doi.org/10.1016/j.jbi.2024.104646.

Malhotra R, Kilic ZS, Martin EB. Artificial intelligence and generative models for materials discovery: a review. arXiv. 2025:arXiv:2508.03278.
https://doi.org/10.48550/arXiv.2508.03278.

Chen Z, Ye W, Chen C, Zuo Y, Ong SP, Li X. A critical analysis of chemical and electrochemical oxidation mechanisms in Li-ion batteries. J Phys Chem C. 2021;125(21):11494-506.
https://doi.org/10.1021/acs.jpcc.1c02553.

Kearns M, Roth A. The ethical algorithm: the science of socially aware algorithm design. Oxford: Oxford University Press; 2020. p. 1-248.

Bolón-Canedo V, Morán-Fernández L, Cancela B, Alonso-Betanzos A. A review of green artificial intelligence: towards a more sustainable future. Neurocomputing. 2024;599:128096.
https://doi.org/10.1016/j.neucom.2024.128096.

Nguyen CT. Trust as an unquestioning attitude. Oxf Stud Epistemol. 2022;7:214-44.

Jo ES, Gebru T. Lessons from archives: strategies for collecting sociocultural data in machine learning. In: Proc Conf Fairness Account Transpar. 2020. p. 306-16.
https://doi.org/10.1145/3351095.3372829.