ORIGINAL RESEARCH **Open access**

Human Oversight as System Design: A Conceptual Reframing of Control in Semi-Autonomous Materials AI

Ahmed El-Kholy[1*], Nour Abdelrahman[1], Karim Hassan[2], Mona Saad[1]

Abstract

The integration of artificial intelligence (AI) into materials science has ushered in an era of semi-autonomous systems that accelerate discovery through predictive modeling, high-throughput screening, and adaptive experimentation. These systems offer substantial promise for addressing global challenges in energy, sustainability, and advanced manufacturing; however, their reliance on data-driven inference introduces risks related to bias propagation, epistemic uncertainty, and misalignment with scientific values. Conventional approaches treat human oversight primarily as an external corrective mechanism—post hoc monitoring or intervention in response to model outputs. This paper proposes a conceptual reframing wherein human oversight is repositioned as an intrinsic element of system design. Rather than viewing control as supervision layered atop an autonomous core, oversight is conceptualized as deliberate architectural choices that embed human judgment into the foundational structure of semi-autonomous materials AI. Drawing on literature from materials informatics, data bias mitigation, explainable AI, and human-AI collaboration, the proposed framework delineates three interdependent dimensions: epistemic boundary-setting, value-aligned modulation, and adaptive reflexivity. This reframing shifts the discourse from mitigating human absence to engineering human presence, fostering systems that are inherently more robust, interpretable, and aligned with the normative goals of scientific inquiry. By reconceptualizing oversight as design, the framework offers a pathway to responsible integration of AI in materials discovery without presupposing full autonomy or diminishing human agency.

Keywords Materials informatics, Data bias, Explainable AI, Human-AI collaboration, Semi-autonomous systems, Oversight design

*Correspondence:
Ahmed El-Kholy
ahmed.elkholy@outlook.com
[1] Department of Materials Engineering and Artificial Intelligence, Faculty of Engineering, Alexandria University, Alexandria, Egypt
[2] Department of Computational Materials Systems, Faculty of Engineering, Ain Shams University, Cairo, Egypt

Introduction

Materials science increasingly operates at the intersection of urgent societal imperatives and rapidly advancing computational capabilities. The development of next-generation materials for sustainable energy storage, carbon capture, quantum technologies, and resilient infrastructure demands exploration across immense chemical and structural design spaces that far exceed the limits of traditional trial-and-error experimentation [1, 2]. As experimental throughput struggles to scale with this combinatorial complexity, materials informatics—rooted in machine learning (ML) and artificial intelligence (AI)—has emerged as a critical paradigm shift. By enabling data-driven modeling of structure–property relationships, inverse design strategies, and closed-loop experimentation, AI systems promise to alter the tempo and scope of materials discovery fundamentally [2, 3].

Within this evolving landscape, semi-autonomous materials AI systems have become increasingly prominent. These systems operate in a hybrid regime in which algorithmic components conduct high-volume tasks—such as candidate screening, surrogate modeling, or experiment selection—while human experts retain authority over pivotal scientific judgments, including hypothesis formulation, validation decisions, and ethical appraisal [4, 5]. Such configurations are often portrayed as a pragmatic compromise between full automation and traditional human-centric research, offering the prospect of compressing decades of iterative discovery into substantially shorter development cycles [6]. However, this acceleration is neither automatic nor guaranteed. System performance remains critically dependent on the quality, diversity, and epistemic adequacy of underlying data, the interpretability of model inferences, and their alignment with broader scientific and societal objectives [2].

A central obstacle arises from the structural constraints of data-driven materials research itself. Materials datasets are frequently sparse, fragmented, and heterogeneous, reflecting the high cost of experimentation, domain-specific measurement practices, and historically contingent research priorities [7, 8]. These limitations introduce systematic selection biases that are readily absorbed by learning algorithms. Rather than facilitating genuine exploration, models may inadvertently reinforce dominant paradigms, privileging well-studied material classes while marginalizing unconventional or high-risk regions of chemical space [9]. Compounding this issue, many state-of-the-art models operate as opaque black boxes, delivering accurate predictions without providing the mechanistic insight required for theory formation, causal reasoning, or scientific generalization [10, 11]. In such settings, predictive success alone is insufficient to sustain long-term scientific progress.

In response to these challenges, a growing body of literature emphasizes the need for human involvement in AI-driven materials workflows. Human-in-the-loop paradigms seek to integrate expert judgment into model training, evaluation, and deployment, allowing researchers to correct anomalies, inject domain knowledge, or redirect exploration trajectories [12, 13]. Self-driving laboratories exemplify this trend by coupling robotic experimentation with AI-based decision engines and intermittent human guidance, thereby closing feedback loops between prediction and physical realization [14, 15]. Across these systems, oversight mechanisms vary in intensity—from active intervention during algorithmic decision-making to post hoc review of outputs and recommendations [16].

Despite these advances, prevailing conceptualizations of human oversight remain largely reactive. Oversight is typically framed as an external supervisory layer—a safeguard intended to monitor, validate, or override algorithmic behavior once it has already been instantiated [17]. While such approaches mitigate certain risks, they implicitly treat human agency as ancillary rather than constitutive. This framing underestimates the extent to which early design choices—regarding objectives, constraints, interaction protocols, and interpretive affordances—shape the epistemic trajectory of semi-autonomous systems long before any supervisory intervention occurs.

This manuscript advances a reframing of human oversight in materials AI, shifting it from a predominantly supervisory function to a foundational design principle. Rather than acting as a corrective afterthought, oversight is conceptualized here as an intentional architecture embedded within the system itself—defining how humans and algorithms interact, where decision boundaries are drawn, and which values govern exploration and optimization [18, 19]. In this view, control is not exercised solely through intervention but through the deliberate configuration of system affordances, feedback pathways, and epistemic constraints. Such an approach resonates with emerging calls for responsible AI in scientific domains, which emphasize transparency, epistemic humility, and alignment with societal benefit as core design commitments rather than external compliance requirements [20, 21].

The remainder of this manuscript synthesizes relevant literature on semi-autonomous materials discovery, human–AI interaction, and responsible scientific AI, and introduces a novel conceptual framework. This framework positions oversight as constitutive of system integrity, offering a theoretical scaffold for designing materials AI systems in which human agency is not peripheral but integral to discovery itself.

Theoretical Background and Literature Synthesis

Materials informatics and the rise of semi-autonomous discovery

Materials informatics has emerged as a foundational approach for addressing the scale and complexity of modern materials discovery. By integrating computational modeling, structured data infrastructures, and machine learning (ML), the field seeks to accelerate the identification, characterization, and optimization of materials beyond what is feasible through conventional experimentation alone [5, 22]. Early efforts emphasized handcrafted descriptors, physics-informed features, and high-throughput screening pipelines that mapped compositional or structural inputs to targeted properties. While effective within constrained domains, these approaches were limited in their ability to generalize across diverse material classes or capture higher-order relational structure.

Recent advances have substantially broadened the methodological repertoire of materials informatics. Graph neural networks enable representations that more faithfully encode atomic connectivity and local environments, while generative models support inverse design by proposing candidate structures conditioned on desired properties [23]. Active learning and Bayesian optimization further extend these capabilities by dynamically selecting experiments or simulations that maximize information gain under resource constraints [24]. Together, these developments have laid the groundwork for semi-autonomous discovery systems that integrate prediction, decision-making, and experimentation into cohesive workflows.

Semi-autonomous materials AI systems operationalize this integration through closed-loop architectures in which algorithms propose experiments, evaluate outcomes, and iteratively refine models with varying degrees of human involvement [25]. Rather than replacing human scientists, such systems redistribute cognitive labor: algorithms manage combinatorial exploration and optimization, while humans define scientific objectives, contextualize results, and interpret emergent patterns [10]. This hybrid configuration reflects an implicit recognition that full autonomy remains unattainable in materials science, where discovery is shaped not only by statistical regularities but also by thermodynamic constraints, synthesizability, scalability, and performance under real-world operating conditions [9]. As a result, semi-autonomy has become the dominant paradigm, foregrounding interaction between human judgment and algorithmic inference.

Data bias and its implications for model reliability

Despite methodological advances, the reliability of materials AI systems remains fundamentally constrained by the characteristics of available data. Materials datasets are often uneven, reflecting historical research priorities, experimental feasibility, and publication incentives rather than systematic coverage of chemical or structural space [7, 8]. Certain material families, synthesis routes, or property regimes are disproportionately represented, while others—such as metastable phases, defective structures, or negative results—remain sparsely documented. For example, widely used crystal structure repositories are biased toward thermodynamically stable compounds, limiting model exposure to transient or kinetically accessible states that are critical for many functional applications [9].

These structural biases have direct consequences for model behavior. Empirical studies demonstrate that models trained on imbalanced datasets may achieve high apparent accuracy while learning spurious correlations that fail to transfer beyond the training distribution [8, 9]. In out-of-distribution regimes—precisely where discovery value is highest—such models often exhibit degraded performance and overconfident predictions. While mitigation strategies such as data augmentation, uncertainty quantification, and algorithmic debiasing have been proposed, these interventions are frequently applied after model development, addressing symptoms rather than underlying causes [10].

The persistence of data bias highlights a deeper issue: design decisions regarding data inclusion, representation, and curation are rarely treated as sites of epistemic governance. Instead, they are framed as technical preprocessing steps, obscuring their role in shaping model assumptions and exploratory trajectories. This gap underscores the need for oversight mechanisms that interrogate data provenance, representational coverage, and implicit value judgments from the earliest stages of system design, rather than relying solely on downstream correction [11].

Human–AI collaboration and oversight mechanisms

Human–AI collaboration has become a central theme in contemporary materials research, encompassing interactive modeling, expert-guided Bayesian optimization, and hybrid autonomous workflows [12–14]. Human-in-the-loop approaches are motivated by the recognition that certain forms of knowledge—such as physical plausibility, experimental practicality, or strategic research priorities—are difficult to infer from data alone [15]. By incorporating expert feedback, these systems can improve predictive performance, avoid infeasible recommendations, and align exploration with contextual goals.

In self-driving laboratories, human oversight plays a critical role in responding to unexpected experimental outcomes, refining hypotheses, and ensuring coherence with long-term research objectives [14]. Oversight mechanisms span a spectrum of engagement, from real-time intervention in decision-making processes to periodic review and validation of system outputs [16]. Across these implementations, a consistent emphasis is placed on preserving human authority in decisions that carry epistemic, ethical, or resource-allocation implications.

However, much of the existing discourse conceptualizes human involvement primarily as a corrective function. Humans are positioned as monitors who validate, override, or adjust algorithmic outputs after they have been generated, rather than as co-designers who shape the structure and constraints of the system itself [17]. This framing implicitly treats oversight as external to core system logic, limiting its capacity to influence upstream assumptions, interaction protocols, and value trade-offs that govern behavior over time.

Epistemic values and responsible innovation in AI-Driven science

Scientific practice is guided not only by technical performance but also by epistemic values such as accuracy, robustness, explanatory adequacy, and generalizability, alongside normative commitments to transparency, equity, and societal benefit [18–20]. In AI-driven materials discovery, these values intersect with growing concerns over explainability, fairness, and accountability [21]. Explainable AI methods—including feature attribution, attention mechanisms, and surrogate models—seek to render algorithmic inferences intelligible, allowing scientists to assess whether predictions are consistent with physical understanding rather than merely statistically successful [10, 11].

Responsible AI frameworks further argue that ethical and epistemic considerations should be embedded early in system design, emphasizing anticipatory governance, bias awareness, and value alignment [20]. Yet applications of these principles within materials science remain underdeveloped. Existing efforts are often fragmented, addressing interpretability, bias, or ethics in isolation rather than as interdependent design dimensions. Moreover, engagement with perspectives from the philosophy of science or ethics remains limited, despite their relevance to questions of explanation, trust, and scientific responsibility [21]. **Table 1** summarizes how prevailing approaches to human oversight in materials AI conceptualize control primarily as external supervision, highlighting the resulting epistemic and normative limitations that motivate the proposed reframing of oversight as system design.

Table 1. Conceptual comparison of oversight paradigms in semi-autonomous materials AI

Dimension of comparison	Conventional oversight paradigm	Oversight-as-design paradigm (this work)
Conceptual role of humans	External supervisors validating or correcting outputs	Intrinsic system designers shaping architecture and constraints
Temporal locus of oversight	Post hoc or episodic intervention	Continuous and anticipatory embedding across the system lifecycle
Treatment of data bias	Detected and mitigated after model training	Prevented through epistemic boundary-setting at the design stage

Handling of uncertainty	Quantified at the prediction level	Governed through domain-scoped inference boundaries
Value integration	External ethical guidelines or audits	Operationalized through multi-objective system design
Human–AI relationship	Monitoring and override	Co-evolution and distributed agency
Failure response	Reactive correction	Proactive architectural governance
Epistemic orientation	Performance-centric	Epistemic integrity–centric

Taken together, the literature reveals a persistent conceptual gap. While human involvement is widely acknowledged as necessary, oversight is rarely theorized as a proactive design element capable of shaping epistemic trajectories and preventing normative failure modes before they arise. This gap motivates the framework proposed in the following section, which reconceptualizes oversight as constitutive of semi-autonomous materials AI systems rather than an auxiliary safeguard.

Proposed conceptual framework: Oversight as system design in semi-autonomous materials AI

The proposed conceptual framework reframes human oversight in semi-autonomous materials AI systems, shifting it from an external supervisory function to an intrinsic design principle. Rather than treating oversight as a downstream corrective mechanism applied after algorithmic outputs are generated, the framework conceptualizes oversight as constitutive of system architecture itself. Human agency is thus embedded within the structural, normative, and adaptive components that govern system behavior over time. This design-oriented perspective recognizes that epistemic risks, value misalignments, and exploratory path dependencies are often introduced upstream, through early architectural and objective-setting decisions, rather than at execution.

The framework is organized around three interdependent dimensions—epistemic boundary-setting, value-aligned modulation, and adaptive reflexivity—that together define a coherent control architecture for semi-autonomous materials-discovery systems. These dimensions do not operate independently; instead, they interact dynamically, distributing human agency across multiple layers of system operation.

Epistemic boundary-setting

Epistemic boundary-setting defines the permissible scope of AI autonomy by delineating where algorithmic inference is considered valid, reliable, and scientifically meaningful. In materials science, where extrapolation beyond observed data can lead to physically implausible or experimentally infeasible predictions, such boundary-setting is critical. Within the proposed framework, humans actively design these boundaries through formal mechanisms such as domain ontologies, uncertainty thresholds, validity criteria, and exclusion rules that constrain inference to epistemically defensible regimes [1, 2].

Crucially, these boundaries are not conceived as rigid constraints but as configurable and revisable design elements. As experimental evidence accumulates and domain understanding evolves, boundary conditions can be systematically expanded, tightened, or restructured. This dynamic configurability allows semi-autonomous systems to balance exploratory ambition with epistemic caution, preventing premature generalization while still enabling progressive knowledge extension. By embedding epistemic limits directly into system logic, boundary-setting transforms oversight from reactive error correction into anticipatory governance of inference behavior.

Value-aligned modulation

While epistemic boundaries regulate where AI systems may operate, value-aligned modulation governs *how* they optimize within those boundaries. This dimension integrates both epistemic and normative values into objective functions, decision protocols, and optimization strategies. Rather than privileging predictive accuracy alone, humans specify multi-objective reward structures that explicitly encode trade-offs among accuracy, interpretability, robustness, resource efficiency, and broader societal priorities such as sustainability or environmental impact [2, 3].

Value-aligned modulation acknowledges that materials discovery is inherently value-laden: choices about which properties to optimize, which candidates to prioritize, and which risks to tolerate are not purely technical decisions. Within this framework, these choices are operationalized through tunable weighting parameters, constraint penalties, and explicit veto points that allow human judgment to override algorithmic recommendations. Such mechanisms allow value commitments to be expressed not as abstract ethical guidelines but as actionable system controls that shape search trajectories and decision outcomes.

Importantly, modulation is designed to be context-sensitive. Value weightings may shift depending on application domain, stage of discovery, or external constraints, enabling adaptive prioritization without sacrificing transparency. In this way, value alignment becomes a continuous, negotiated process embedded in system operation rather than a one-time specification.

Adaptive reflexivity

The third dimension, adaptive reflexivity, addresses the temporal evolution of oversight itself. Semi-autonomous materials AI systems operate in environments characterized by uncertainty, shifting scientific paradigms, and incomplete knowledge. Adaptive reflexivity enables continuous co-evolution between human and AI components by embedding reflective feedback loops that operate at a meta-level. These loops incorporate human meta-cognition—critical reflection on model assumptions, detection of emergent biases, reassessment of explanatory adequacy, and recognition of paradigm shifts—into system updates and reconfiguration processes [4, 5].

Unlike conventional feedback loops that focus on performance metrics or prediction error, reflexive loops target the underlying premises of system operation. They allow humans to question not only whether the system is performing well, but whether it is asking the right questions, privileging the right representations, and pursuing scientifically meaningful objectives. Over time, this enables a transition from tactical oversight—intervening in isolated decisions—to strategic oversight, in which humans reshape system structure, objectives, and interaction protocols in response to accumulated insight.

Adaptive reflexivity thus ensures that oversight itself remains responsive to evolving scientific understanding, preventing ossification of assumptions and supporting long-term epistemic resilience.

Integrated control architecture

Taken together, these three dimensions form an integrated control architecture in which human agency is distributed across layers rather than centralized at a single point of intervention. Epistemic boundary-setting constrains the domain of action, value-aligned modulation shapes optimization behavior within that domain, and adaptive reflexivity governs system evolution over time. Oversight, in this formulation, is neither episodic nor adversarial; it is structural, continuous, and generative.

By embedding human judgment into the foundational design of semi-autonomous materials AI systems, the framework offers a theoretical scaffold for aligning acceleration with epistemic integrity and societal responsibility. The next section elaborates on the analytical implications of this architecture for materials discovery workflows. It outlines how oversight-as-design can mitigate epistemic and normative failure modes inherent in autonomous exploration. The integrated scaffolding framework **(Figure 1)** positions human designers as external nodes interfacing with epistemic, value-aligned, and reflexive oversight layers.

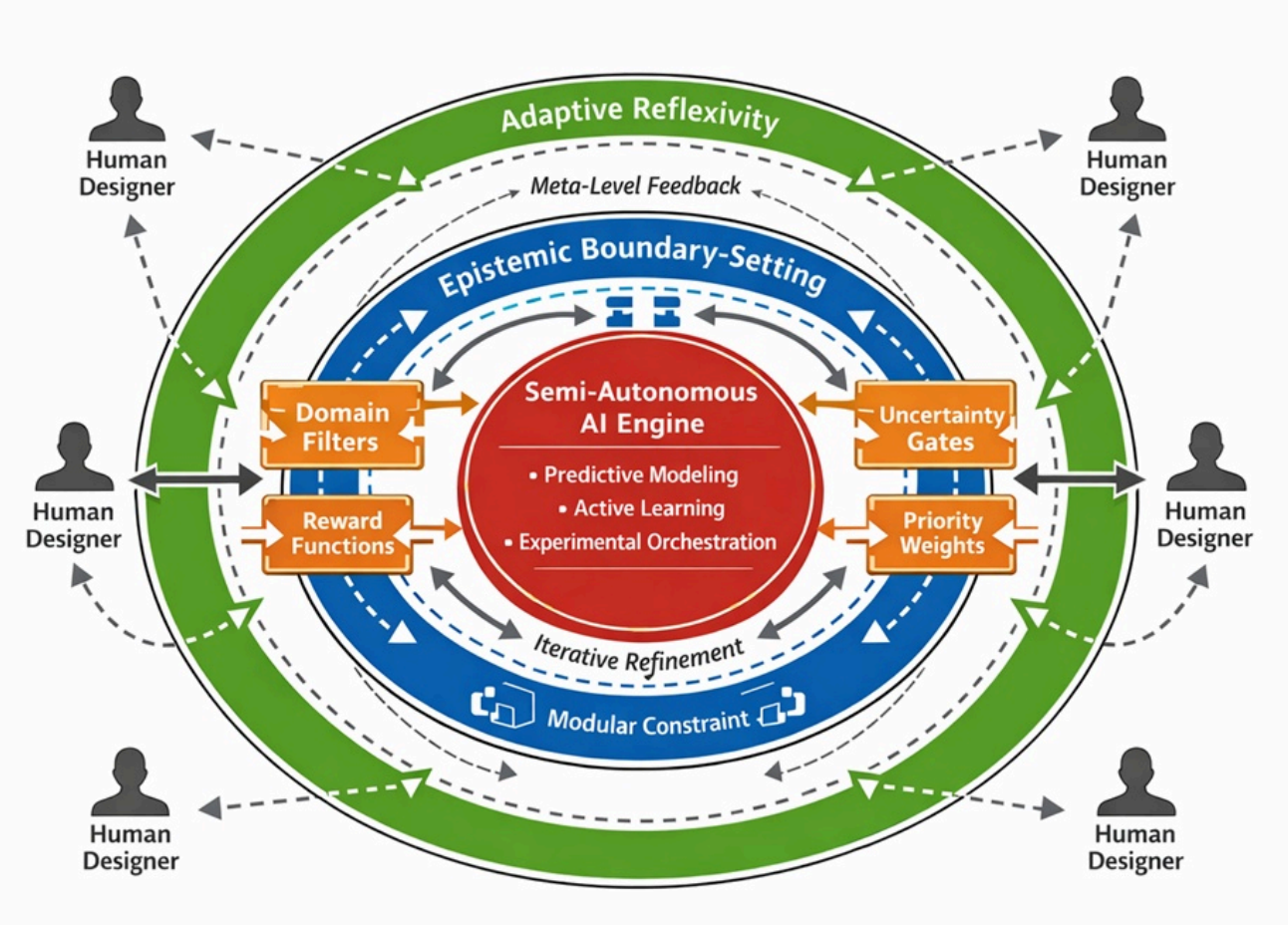

Figure 1. The integrated scaffolding framework for proactive AI oversight: A multi-layered architecture embedding epistemic, value-aligned, and reflexive human-AI collaboration.

Propositions

The conceptual framework advanced in this manuscript yields a set of theoretical propositions that articulate the implications of reframing human oversight as an intrinsic design element in semi-autonomous materials AI systems. Collectively, these propositions specify how architectural embedding of human agency reshapes robustness, alignment, and long-term epistemic reliability.

Proposition 1

Epistemic boundary-setting, when architected as an intrinsic design layer, enhances the robustness of semi-autonomous materials AI by constraining inference to domains of validated physicochemical plausibility, thereby reducing the propagation of extrapolative errors characteristic of data-driven models [5, 8, 9, 22].

This proposition asserts that formal epistemic boundaries—such as ontology-constrained search spaces, uncertainty-calibrated validity regions, and exclusion criteria—operate not as restrictive barriers but as enabling structures. By embedding these constraints at the architectural level, the system internalizes epistemic humility, prioritizing scientific credibility over unconstrained predictive reach while preserving meaningful exploratory capacity.

Proposition 2

Value-aligned modulation, implemented through multi-objective design protocols, ensures that semi-autonomous materials AI systems reflect pluralistic scientific and societal values—such as sustainability, synthesizability, and

interpretability—without reliance on post hoc correction [2, 3, 18–20].

Here, modulation is operationalized through tunable objective functions, weighted trade-offs, and explicit human-defined veto points. Embedding value alignment within system logic transforms ethical and normative considerations from external overlays into operational principles, enabling proactive navigation of trade-offs inherent in materials discovery.

Proposition 3

Adaptive reflexivity, realized through meta-level feedback loops, enables the co-evolution of human judgment and AI capabilities, transforming semi-autonomous materials AI into systems capable of paradigm-level adaptation in response to emergent scientific insight [4, 5, 14, 15].

This proposition posits that reflexivity extends oversight beyond performance tuning toward strategic reconfiguration. By incorporating human meta-cognition—critical reflection on assumptions, bias recognition, and theoretical reassessment—into iterative system updates, oversight evolves from tactical intervention to long-horizon epistemic stewardship.

Proposition 4

The interdependence of epistemic boundary-setting, value-aligned modulation, and adaptive reflexivity generates a synergistic control architecture in which omission of any single dimension undermines overall system integrity, necessitating holistic rather than piecemeal design [1, 2, 10, 11].

Interdependencies manifest through cascading effects: weak epistemic boundaries amplify value misalignment, insufficient reflexivity impedes boundary recalibration, and unmodulated objectives exacerbate epistemic drift. This proposition emphasizes that oversight effectiveness emerges from architectural coherence rather than isolated supervisory mechanisms.

Proposition 5

By repositioning human oversight as constitutive system design rather than external supervision, semi-autonomous materials AI transitions from mitigation-centric paradigms to presence-centric paradigms, thereby enhancing long-term epistemic and normative reliability [12, 13, 17, 21].

Table 2 consolidates the proposed framework by mapping each oversight dimension to its corresponding theoretical propositions, control mechanisms, and anticipated epistemic outcomes, clarifying how oversight-as-design operates as an integrated system architecture.

Table 2. Oversight-as-design framework: dimensions, propositions, and system-level functions

Oversight dimension	Core design function	Associated propositions	Primary system-level outcome
Epistemic boundary-setting	Constrains inference to validated physicochemical regimes through ontologies, uncertainty thresholds, and exclusion rules	Proposition 1, Proposition 4	Enhanced robustness; reduced extrapolative error; epistemic humility
Value-aligned modulation	Embeds pluralistic scientific and societal values into objective functions and decision protocols	Proposition 2, Proposition 4	Proactive value alignment; transparent trade-off navigation
Adaptive Reflexivity	Enables meta-level feedback, incorporating human reflection into system reconfiguration	Proposition 3, Proposition 4	Long-term adaptability; paradigm-level learning

Integrated Control Architecture	Distributes human agency across epistemic, normative, and temporal layers	Proposition 5	Sustained epistemic and normative reliability

This synthesizing proposition captures the framework's core reframing: human agency is not a compensatory feature but an engineered presence distributed across system layers, yielding AI systems that are more transparent, adaptable, and scientifically aligned.

Results and Discussion

The proposed propositions collectively signal a conceptual shift in how control and responsibility are theorized within semi-autonomous materials AI. Conventional paradigms predominantly treat human involvement as a safeguard against algorithmic limitation—monitoring outputs, intervening in anomalous cases, or validating predictions after generation [16, 17]. While such approaches reduce immediate risk, they position oversight as supplementary rather than foundational, leaving core system dynamics shaped primarily by algorithmic optimization.

In contrast, the present framework embeds oversight through deliberate architectural choices spanning epistemic, normative, and reflexive dimensions. This reframing aligns with emerging trends in materials informatics toward hybrid intelligence, where human domain expertise complements data-driven inference rather than merely correcting it [12–15]. Epistemic boundary-setting directly addresses persistent challenges of data bias and poor generalization by constraining inference to regimes supported by physicochemical reasoning, resonating with calls for uncertainty-aware and domain-informed modeling [7–10].

Value-aligned modulation responds to growing recognition that materials discovery is inherently value-laden, involving trade-offs among performance, sustainability, interpretability, and social impact [18–21]. By embedding multi-objective optimization and tunable priorities into the system logic, the framework operationalizes responsible AI principles directly, reducing reliance on external ethical audits or retrospective evaluation.

Adaptive reflexivity addresses the temporal dynamics of scientific progress, in which theoretical shifts, anomalous results, or evolving societal priorities necessitate the reassessment of both models and objectives. Drawing on human–AI collaboration literature, reflexivity elevates oversight from episodic correction to strategic evolution, supporting long-term epistemic resilience [4, 5, 14, 15].

Potential limitations of the framework include challenges in operationalizing reflexivity at scale, given the cognitive and institutional costs of sustained human meta-reflection, as well as the risk that overly restrictive boundaries could suppress serendipitous discovery. Future theoretical work may formalize inter-layer dependencies, develop metrics for architectural coherence, or explore mechanisms for balancing constraint with creative exploration.

Conclusion

This manuscript advances a conceptual reframing in which human oversight is understood not as external supervision but as an intrinsic dimension of system design in semi-autonomous materials AI. By articulating epistemic boundary-setting, value-aligned modulation, and adaptive reflexivity as interdependent architectural elements, the framework relocates control from reactive intervention to foundational engineering.

The derived propositions clarify how such embedding enhances robustness, alignment, and long-term adaptability, offering a principled alternative to mitigation-centric oversight models. As materials informatics continues to accelerate discovery, this perspective provides a theoretical scaffold for designing semi-autonomous systems that balance speed with epistemic integrity and normative responsibility. Future work may translate this framework into formal models,

implementation guidelines, or interdisciplinary extensions, contributing to the responsible evolution of AI-driven materials science.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 09 Apr 2025 Revised: 04 May 2025 Accepted: 04 Jul 2025
Published online: 18 January 2026

References

Xu P, Ji X, Li M, Lu W. Small data machine learning in materials science. npj Comput Mater. 2023;9(1):42.

Morgan D, Jacobs R. Opportunities and challenges for machine learning in materials science. Annu Rev Mater Res. 2020;50:71-103.

Fujii M. Significance of materials informatics and the development of new materials. JSAP Rev. 2022;2022:220416.

Noack MM, Doerk GS, Li R, Streit JK, Vaia RA, Yager KG, et al. Autonomous materials discovery driven by gaussian process regression. Sci Rep. 2020;10:17663.

Li C, Zheng K. Methods, progresses, and opportunities of materials informatics. InfoMat. 2023;5(6):e12425.

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

Zhang H, Chen WW, Rondinelli JM, Chen W. ET-AL: Entropy-targeted active learning for bias mitigation in materials data. arXiv. 2022;2211.07881.

Kumagai M, Ando Y, Tanaka A, Tsuda K, Katsura Y, Kurosaki K. Effects of data bias on machine-learning–based material discovery using experimental property data. Adv Mater Interfaces. 2022;9:2109447.

Zhang H, Chen W, Rondinelli JM, Chen W. Mitigating bias in scientific data: A materials science case study. NeurIPS. 2023.

Oviedo F, et al. Explainable machine learning in materials science. npj Comput Mater. 2022;8:184.

Lu Z, Chen X, Liu X, Lin D, Wu Y, Zhang Y, et al. Interpretable machine-learning strategy for soft-magnetic property. npj Comput Mater. 2020;6:187.

Wang HC, Schmidt J, Marques MA, Wirtz L, Romero AH. Symmetry-based computational search for novel binary and ternary 2d materials. 2D Mater. 2023;10(3):035007.

Chen D, Bai Y, Ament S, Zhao W, Guevarra D, Zhou L, et al. Automating crystal-structure phase mapping by combining deep learning with constraint reasoning. Nat Mach Intell. 2021;3(9):812-22.

Koscher BA, et al. Autonomous, multiproperty-driven molecular discovery. Science. 2023;382(6672):eadl1407.

Hysmith H, Foadian E, Padhy SP, Kalinin SV, Moore RG, Ovchinnikova OS, et al. The future of self-driving laboratories: From human in the loop interactive ai to gamification. Digit Discov. 2024;3(4):456-68.
https://doi.org/10.1039/D4DD00040D.

Tobias AV, Wahab A. Autonomous ‘self-driving’ laboratories: A review of technology and policy implications. R Soc Open Sci. 2025;12(7):250646.
https://doi.org/10.1098/rsos.250646.

Hung L, Yager JA, Monteverde D, Baiocchi D, Kwon H-K, Sun S, et al. Autonomous laboratories for accelerated materials discovery: A community survey and practical insights. Digit Discov. 2024;3:1273-9.

Responsible ai in materials science contexts. Various sources. 2023-2024.

Ethical considerations in ai-driven scientific discovery. 2024.

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.

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

Chelladurai U, Pandian S. A novel blockchain based electronic health record automation system for healthcare. J Ambient Intell Humaniz Comput. 2022;13(1):693-703.

Wang Z, Chen A, Tao K, Cai J, Han Y, Gao J, et al. Alphamat: A material informatics hub connecting data, features, models and applications. npj Comput Mater. 2023;9(1):130.

Katsura Y, Akiyama M, Morito H, Fujioka M, Sugahara T. Systematic searches for new inorganic materials assisted by materials informatics. Sci Technol Adv Mater. 2024;26(1):2428154.
https://doi.org/10.1080/14686996.2024.2428154.

Bayley O, Savino E, Slattery A, Noel T. Autonomous chemistry: Navigating self-driving labs in chemical and material sciences. Matter. 2024;7(7):2382-98.