REVIEW **Open access**

Decision Authority Frameworks in Autonomous Materials Discovery Systems

Ravi Kumar[1*], Neha Sharma[1], Arjun Nair[2]

Abstract

The rapid evolution of computational and data-driven materials engineering has ushered in autonomous discovery systems that integrate machine learning, high-throughput simulations, and robotic experimentation to accelerate materials innovation. Central to these systems are decision authority frameworks, which define how authority is delegated between human operators and artificial intelligence agents, ensuring safe, ethical, and efficient operations. This review synthesizes recent literature on delegation models, human override mechanisms, responsibility assignment, and policy encoding within materials informatics ecosystems. We examine how these frameworks operate in closed-loop discovery pipelines, where active learning and uncertainty quantification guide iterative experimentation. Key areas include representation learning via graph neural networks for materials property prediction, multimodal dataset integration for simulation-experiment synergy, and inverse design strategies that balance exploration and exploitation. By analyzing delegation in autonomous laboratories, we highlight the role of human-in-the-loop paradigms in mitigating risks such as algorithmic bias or experimental failures. The review underscores the need for robust policy encodings that embed ethical constraints and regulatory compliance into AI-driven workflows. Drawing from high-impact studies, we provide an integrative perspective on how these frameworks enhance reliability in materials discovery, paving the way for scalable, trustworthy autonomous systems in computational materials science.

Keywords Autonomous materials discovery, Decision authority, Delegation models, Human override, Responsibility assignment, Policy encoding

*Correspondence:
Ravi Kumar
ravi.kumar@outlook.com
[1] Department of Materials Data Engineering, Faculty of Engineering, IIT Delhi, New Delhi, India
[2] Department of Computational Materials Systems, Faculty of Engineering, IIT Bombay, Mumbai, India

Introduction

The field of computational and data-driven materials engineering has transformed from traditional trial-and-error approaches to sophisticated, AI-assisted paradigms that leverage vast datasets and predictive modeling to expedite the discovery of novel materials. At the heart of this transformation lie autonomous materials discovery systems, which orchestrate end-to-end workflows encompassing data acquisition, model training, hypothesis generation, and experimental validation [1-3]. These systems, often termed self-driving laboratories, integrate computational tools such as density functional theory simulations with robotic platforms for high-throughput synthesis and characterization, enabling closed-loop optimization that minimizes human intervention [4-6]. However, the increasing autonomy of these systems raises critical questions about decision-making authority: who—or what—holds the reins in critical junctures, and how is accountability ensured?

Decision authority frameworks in autonomous materials discovery systems provide structured mechanisms for allocating control between AI agents and human experts.

These frameworks encompass delegation models, which specify the scope of AI autonomy; human override rights, allowing intervention in anomalous scenarios; responsibility assignment, delineating liability for outcomes; and policy encoding, embedding rules and constraints into the system's architecture [7-9]. In materials engineering contexts, such frameworks are essential for navigating the complexities of data-driven ecosystems, where uncertainties in predictions can lead to inefficient resource allocation or safety hazards [10, 11]. For instance, in active learning systems, AI might autonomously select the next experiment based on uncertainty quantification, but human overrides are crucial when dealing with hazardous materials or unmodeled phenomena [8].

The origins of these frameworks can be traced to broader AI governance principles, adapted to the unique demands of materials science. Early work in materials informatics focused on machine learning for property prediction, using techniques like graph neural networks to represent crystal structures and predict bandgaps or mechanical properties [2, 12]. As systems evolved toward autonomy, integration of multimodal datasets—combining simulation outputs, experimental spectra, and literature-mined knowledge—became pivotal [13, 14]. High-throughput computation platforms, such as those employing automated density functional theory workflows, generate terabytes of data, necessitating AI-driven curation and analysis [1, 15]. Yet, without proper authority frameworks, these platforms risk amplifying errors, such as propagating biases from incomplete datasets into discovery loops [7].

Delegation models vary in sophistication, from fully autonomous modes where AI handles routine tasks like parameter optimization, to hybrid models incorporating human feedback loops [5, 8, 16]. In inverse materials design, where target properties guide the search for candidate structures, delegation often involves policy-based reinforcement learning to balance exploration (sampling diverse chemical spaces) and exploitation (refining promising candidates) [2, 3, 17]. Human override rights are particularly salient in autonomous laboratories, where robotic systems execute syntheses; overrides prevent catastrophic failures, such as unintended reactions in flow chemistry setups [4, 18, 19]. Responsibility assignment draws from systems engineering, assigning accountability to AI for computational decisions and to humans for ethical oversight [7, 9, 20]. Policy encoding translates these into computable forms, such as constraint satisfaction problems integrated into optimization algorithms [11, 12].

The relevance of these frameworks is amplified by the interdisciplinary nature of modern materials engineering. For example, in nanomedicine development, self-driving labs optimize nanoparticle formulations through closed-loop iterations, but authority frameworks ensure compliance with biomedical regulations [3, 19]. Similarly, in energy materials discovery, such as perovskites for photovoltaics, AI-driven phase mapping benefits from human-in-the-loop corrections to Bayesian models [8]. Recent community perspectives highlight the need for standardized frameworks to foster collaboration across international materials acceleration platforms [1, 10, 12].

This review positions itself at the intersection of computational workflows and governance in autonomous systems. By synthesizing literature, We offer an original interpretive structure that categorizes decision authority along axes of autonomy level, risk profile, and integration depth. Unlike prior reviews that emphasize technical implementations [1, 2], our focus is on the socio-technical dimensions, providing a systems-level synthesis that informs the design of next-generation discovery platforms. We begin by surveying the landscape of computational and data-driven materials engineering, then delve into autonomous and closed-loop systems, setting the stage for subsequent discussions on challenges and future directions. As mapped in **Figure 1**, decision authority is stratified according to experimental risk, epistemic uncertainty, and regulatory sensitivity.

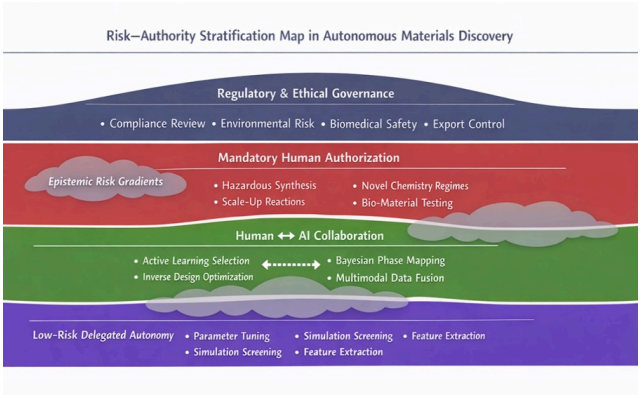

Figure 1. Risk–Authority Stratification Map in Autonomous Materials Discovery

Landscape of Computational & Data-Driven Materials Engineering

The landscape of computational and data-driven materials engineering encompasses a multifaceted ecosystem where advanced algorithms, vast datasets, and integrated workflows converge to drive innovation. At its core, materials informatics serves as the foundational pillar, employing data mining and machine learning to extract patterns from structured repositories like the Materials Project or AFLOW [1, 2]. Graph neural networks have emerged as a dominant tool for representation learning, encoding atomic connectivity and electronic structures into vector embeddings that facilitate property predictions across diverse materials classes, from alloys to polymers [2, 13]. These networks enable efficient navigation of high-dimensional chemical spaces, reducing the computational cost of ab initio simulations [15, 21].

High-throughput computation forms another critical component, automating quantum mechanical calculations to screen thousands of candidates rapidly. Platforms integrating density functional theory with machine learning surrogates accelerate this process, allowing for real-time property assessments in discovery campaigns [1]. For instance, in perovskite design, high-throughput workflows combine simulation data with experimental validation to identify stable compositions [8]. Multimodal datasets further enrich this landscape, fusing disparate sources such as X-ray diffraction patterns, molecular dynamics trajectories, and spectroscopic data [13, 14]. Integration of these datasets via representation learning techniques ensures compatibility, enabling holistic models that capture structure-property relationships [2, 12].

Active learning systems represent a dynamic evolution, where models iteratively query new data points to refine predictions. Uncertainty quantification, often via Bayesian methods or ensemble variance, guides this selection, prioritizing informative samples to optimize resource use [8, 11]. In materials contexts, active learning mitigates the curse of dimensionality in vast search spaces, as seen in polymer blend optimization where evolutionary algorithms formulate blends autonomously [17]. Simulation-experiment integration bridges the gap between virtual and physical realms, with closed-loop systems feeding experimental outcomes back into computational models for recalibration [3, 4, 6]. This synergy is exemplified in nanoparticle synthesis, where robotic platforms adjust parameters based on real-time feedback from machine learning predictors [19].

Inverse materials design flips the traditional paradigm, starting from desired properties to generate candidate structures. Techniques like generative adversarial networks or variational autoencoders, informed by graph representations, produce viable materials blueprints [2, 16]. Within this landscape, decision authority frameworks emerge as essential orchestrators, delegating routine computations to AI while reserving human input for interpretive tasks [5, 7, 8]. For example, in international materials acceleration platforms, brokering mechanisms allocate resources across tenants, with policies encoding fair usage and data sharing [12].

Human-in-the-loop configurations enhance robustness, allowing overrides in cases of model uncertainty or ethical concerns [8, 22]. Responsibility assignment in these ecosystems assigns AI accountability for predictive accuracy, while humans oversee experimental safety [7, 9, 20]. Policy encoding integrates constraints, such as energy efficiency or material sustainability, into optimization objectives [11, 12]. Community surveys underscore the adoption barriers, noting the need for standardized interfaces in autonomous setups [10, 23].

Evolution of delegation models

Delegation models have evolved from static rule-based systems to adaptive AI-driven frameworks. Early models in materials informatics delegated simple tasks like data preprocessing, but modern variants employ reinforcement learning for dynamic authority allocation [2, 5, 16]. In self-driving labs, delegation encompasses experiment planning, with AI selecting synthesis routes based on learned policies [3, 4, 6]. Human overrides are triggered by thresholds in uncertainty metrics, ensuring intervention only when necessary [8].

Integration in materials informatics

In materials informatics, decision frameworks facilitate seamless data flow across pipelines. Graph neural networks, augmented with policy encodings, prioritize features relevant to discovery goals [2, 13]. High-throughput systems benefit from delegated uncertainty quantification, automating batch selections for computation [1, 11].

Challenges in multimodal integration

Multimodal datasets pose integration challenges, addressed by frameworks that delegate fusion tasks to specialized AI modules [13, 14]. Responsibility assignment ensures traceability, mapping errors back to data sources [7].

This landscape illustrates how decision authority frameworks underpin the reliability and scalability of computational materials engineering, fostering a cohesive environment for innovation [1, 10, 12].

Autonomous & closed-loop discovery systems

Autonomous and closed-loop discovery systems epitomize the pinnacle of integration in computational materials engineering, where AI agents orchestrate iterative cycles of hypothesis generation, experimentation, and refinement with minimal human input [1, 2, 4]. These systems embed decision authority frameworks to manage the delegation of tasks, ensuring that AI autonomy aligns with overarching goals and safety protocols [5, 7, 8]. In closed-loop setups, active learning drives the process: models predict outcomes, quantify uncertainties, and select actions that maximize information gain, forming a feedback loop that refines predictions over iterations [8, 11]. **Table 1** summarizes the operational dimensions through which decision authority is encoded, delegated, and governed in autonomous discovery infrastructures.

Table 1. Operational Dimensions of Decision Authority in Autonomous Materials Discovery Systems

Dimension	AI Delegation Role	Human Authority Role	Enabling Technologies
Experiment Selection	Active learning optimization	Override approval	Bayesian optimization, RL
Materials Design	Generative inverse modeling	Ethical review	VAEs, GANs, GNNs
Data Fusion	Multimodal integration	Validation auditing	Representation learning
Robotic Execution	Autonomous synthesis control	Emergency intervention	Self-driving labs
Resource Allocation	Scheduling optimization	Institutional policy setting	Multi-agent orchestration
Risk Assessment	Predictive hazard scoring	Final authorization	Uncertainty quantification

Table 1. Operational dimensions structuring decision authority allocation in autonomous materials discovery systems. The table delineates how AI and human actors share responsibility across experimental planning, execution, and governance layers, mediated by enabling computational infrastructures and policy mechanisms.

Delegation models in these systems range from hierarchical structures, where high-level policies are human-defined and low-level executions are AI-managed, to fully adaptive models using multi-agent architectures [11, 16]. For instance, in chemical self-driving laboratories, orchestration software like ChemOS delegates experiment scheduling to robotic arms while encoding policies for resource constraints [6, 16]. Human override rights are integral, activated via interfaces that monitor system states and allow intervention, such as halting a synthesis if anomalies in real-time data are detected [8, 22]. Responsibility assignment frameworks attribute outcomes: AI bears computational decisions, but humans retain liability for policy violations or ethical lapses [7, 9, 20].

Policy encoding translates abstract rules into actionable constraints, often formalized as optimization problems within the discovery loop. A conceptual formula for this integration can be expressed as:

$$\arg\max_{a \in A} \left[U(\theta \mid a) + \lambda P(a) \right] \quad (1)$$

where a is the selected action (e.g., next experiment), U(θ|a) represents the expected utility in updating model parameters θ, P(a) encodes policy compliance (e.g., safety bounds), and λ balances utility and policy adherence. This formula synthesizes the trade-off in closed-loop systems, delegating maximization to AI while allowing human overrides on λ [5, 11, 12].

In materials discovery, these systems apply to diverse domains. Autonomous phase mapping employs Bayesian

optimization with human-in-the-loop corrections, delegating iterative sampling to AI [8]. In nanomedicine, closed-loops optimize formulations through flow chemistry, with policies encoding biocompatibility checks [3, 4, 18]. Graph neural networks enhance representation in these loops, learning from multimodal data to inform inverse design [2, 13, 14]. Uncertainty quantification ensures robust delegation, flagging high-risk actions for override [8, 11]. **Figure 2** illustrates the orchestration of decision authority across the closed-loop autonomous discovery cycle, highlighting the interaction between AI delegation, policy constraints, and human override mechanisms.

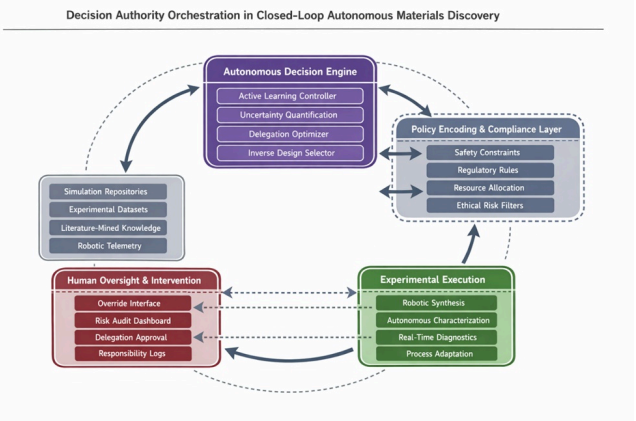

Figure 2. Decision Authority Orchestration in Closed-Loop Autonomous Materials Discovery, The architecture depicts how AI delegation engines coordinate experimental execution under policy constraints while remaining subject to human override and responsibility governance. Multimodal data feedback continuously recalibrates authority allocation across discovery iterations.

Community perspectives emphasize the scalability of these systems, advocating for minimal working examples that incorporate authority frameworks from inception [1, 10, 23]. Multi-agent labs optimize workflows by delegating subtasks, such as data brokering in international platforms [11, 12]. Inverse design in closed-loops benefits from evolutionary optimization, with policies guiding blend formulations [17]. Overall, these systems demonstrate how decision authority frameworks enable efficient, trustworthy autonomy in materials engineering [1, 2, 7].

Challenges & Limitations

Despite rapid progress in decision-authority architectures for autonomous materials discovery, a series of structural, epistemic, operational, and governance-level constraints continue to limit their scalability and reliability across computational and data-driven materials engineering ecosystems [1, 2, 7]. These challenges are not isolated technical deficiencies but emerge from the systemic coupling of machine intelligence, experimental infrastructure, and institutional oversight. As such, limitations manifest simultaneously at the levels of representation, delegation logic, human governance, cyber-physical integration, and regulatory legitimacy.

Epistemic uncertainty and delegation fragility

A foundational limitation resides in how delegation frameworks operationalize uncertainty under conditions of incomplete materials knowledge. Autonomous discovery systems frequently rely on active learning policies in which experimental acquisition is guided by model-estimated uncertainty landscapes. However, these landscapes are themselves conditioned by bounded training distributions and sparse sampling of vast compositional and structural design spaces [8, 11].

When uncertainty estimates are derived from incomplete or density-skewed datasets, delegated decisions may converge toward epistemically misleading regions—areas that appear informative computationally but lack physical relevance. This misalignment is particularly pronounced in early-stage exploration of high-dimensional chemical systems such as high-entropy alloys, hybrid perovskites, or metastable catalytic phases.

Graph neural networks (GNNs), while demonstrating strong interpolation performance within trained chemical manifolds, remain structurally limited in extrapolative reasoning. Their message-passing architectures encode local relational inductive biases that degrade when confronted with novel bonding environments, coordination motifs, or thermodynamic regimes absent from training corpora [2, 13, 21]. In closed-loop settings, such extrapolation failures can propagate through iterative delegation cycles, amplifying experimental misallocation and reinforcing model overconfidence.

The challenge intensifies in multimodal fusion contexts. Autonomous systems increasingly integrate heterogeneous data streams—density functional theory outputs, spectroscopy signatures, microscopy imagery, and robotic synthesis metadata. These modalities operate across divergent noise profiles, spatial resolutions, and epistemic

fidelities. Policy encodings that attempt to harmonize such signals often lack the representational elasticity to reconcile cross-modal inconsistencies, leading to unstable delegation thresholds and conflicting acquisition priorities [12-14].

Human override and governance latency

Human override mechanisms are frequently positioned as ethical and operational safeguards within decision-authority hierarchies. Yet their practical efficacy remains constrained by temporal, cognitive, and interface-design limitations.

In high-throughput self-driving laboratories, robotic execution cycles operate on timescales that significantly outpace human supervisory capacity. Autonomous synthesis platforms can initiate, execute, and analyze experimental runs within minutes, rendering real-time human intervention impractical except under pre-encoded halt conditions [3, 4, 6]. Consequently, override rights function more as retrospective auditing tools than proactive governance instruments.

Interface design further complicates usability. Community assessments indicate that override dashboards often lack intuitive risk visualizations, decision traceability, or uncertainty interpretability layers, resulting in either underutilization or erroneous activations [10, 23]. When human supervisors cannot rapidly interpret why a system selected a given experiment, intervention becomes cognitively burdensome and operationally delayed.

Responsibility attribution introduces an additional governance challenge. Multi-agent discovery ecosystems distribute decision authority across surrogate models, optimization agents, robotic planners, and resource schedulers. Failures rarely originate from a single delegation node; instead, they emerge from cascaded interactions. Without granular logging architectures and lineage-tracking protocols, tracing erroneous outcomes back to specific delegation policies remains technically and legally ambiguous [7, 11, 16].

Policy encoding and computational overhead

Translating ethical, regulatory, and institutional constraints into machine-readable delegation policies constitutes a non-trivial design burden. Materials discovery often intersects with safety-critical domains—energy storage, nanomedicine, catalysis, and environmental remediation—where experimental actions may carry downstream societal risks.

Encoding such constraints requires hybrid expertise spanning materials science, regulatory law, and AI systems engineering. Nuanced directives—such as toxicity thresholds, precursor handling restrictions, or environmental exposure limits—must be formalized into computable policy grammars. These grammars, when embedded into decision loops, introduce additional computational layers that can slow optimization cycles and reduce experimental throughput [9, 12, 20].

Moreover, rigid policy codifications risk brittleness. Overly conservative encodings may suppress exploratory discovery, while permissive policies may fail to prevent hazardous delegations. Achieving adaptive policy elasticity —where governance evolves alongside system learning— remains an unresolved architectural challenge.

Scalability and infrastructure interoperability

As autonomous discovery expands into international materials acceleration platforms, delegation frameworks must operate across distributed infrastructures with heterogeneous governance models.

Cross-institutional collaboration introduces policy conflicts concerning data sovereignty, intellectual property, and experimental prioritization. Delegation brokers tasked with resource allocation must reconcile competing institutional mandates, potentially generating contention over robotic access, compute cycles, or proprietary datasets [16]. Without standardized inter-tenant negotiation protocols, scalability becomes administratively constrained.

Cyber-physical integration presents parallel bottlenecks. Self-driving laboratories rely on tight coupling between algorithmic planning layers and robotic execution systems. Yet hardware variability—pump instabilities in flow chemistry, deposition irregularities in thin-film synthesis, or thermal gradients in furnace systems—introduces stochastic deviations that delegation policies rarely model explicitly [4, 18, 19]. As a result, delegated experiments may fail not due to flawed scientific reasoning but due to unmodeled physical contingencies.

Human-in-the-loop mitigations, while valuable for anomaly detection, reintroduce subjectivity. Operator interventions can bias acquisition trajectories, privileging familiar chemistries or historically successful synthesis routes, thereby constraining algorithmic exploration diversity [5, 8, 22].

Technical Limitions in Simulation–Experiment Integration

Closed-loop coupling between simulation and experiment remains one of the most technically challenging integration frontiers.

Bayesian optimization frameworks, widely used in phase mapping and compositional exploration, assume calibrated posterior uncertainties. In sparse data regimes, however, posterior distributions can become overconfident due to kernel priors or surrogate mis-specification. This leads to premature delegation of high-risk experimental campaigns under the false premise of predictive certainty [1, 8].

Inverse design pipelines face an orthogonal limitation rooted in the “no free lunch” theorem: no single optimization or delegation strategy performs optimally across all materials classes [2, 16, 17]. Representation learning architectures that excel in crystalline solids may fail in polymeric, amorphous, or biomolecular systems. Consequently, delegation policies trained within one materials ontology may degrade when transferred across chemical domains, limiting generalizable autonomy.

Ethical and Regulatory Challenges

Ethical risk intensifies as decision authority shifts from human experts to autonomous systems capable of initiating physical experiments.

In nanomedicine and bio-materials design, optimization objectives—such as maximizing cellular uptake or catalytic reactivity—may inadvertently converge on toxic or environmentally persistent formulations if not bounded by robust safety policies [3, 7]. Autonomous escalation, wherein one delegated decision triggers downstream high-impact experiments, mirrors broader AI alignment concerns but carries tangible material consequences.

Regulatory frameworks remain underdeveloped for such scenarios. Questions of liability—whether borne by algorithm designers, laboratory operators, or institutional owners—lack legal precedent. This governance vacuum slows adoption in safety-critical sectors despite technological readiness [7, 9].

Adoption Barriers and Sociotechnical Friction

Beyond technical constraints, sociotechnical adoption barriers impede widespread deployment.

A lack of standardized architectural frameworks fragments development efforts across laboratories and consortia. Community studies highlight the absence of interoperable delegation taxonomies, benchmarking protocols, and validation metrics, making cross-platform reproducibility difficult [1, 10].

Minimal working examples—often showcased as proof-of-concept self-driving labs—demonstrate feasibility but fail to scale across diverse materials ecosystems with varying synthesis modalities, characterization pipelines, and data ontologies [23].

Finally, a persistent curiosity–creativity gap limits AI’s exploratory agency. While autonomous systems excel at optimizing within defined objective manifolds, they struggle to formulate radically novel hypotheses, unconventional synthesis pathways, or paradigm-shifting materials concepts. Delegation thus remains bounded by algorithmically encoded imagination rather than open-ended scientific intuition [9, 20].

Collectively, these challenges reveal that decision-authority frameworks are constrained not merely by algorithmic maturity but by deeper epistemic, infrastructural, and governance entanglements. Addressing uncertainty calibration, override latency, policy computability, infrastructure interoperability, and ethical accountability will be essential for transitioning autonomous materials discovery from experimental novelty to institutionalized scientific infrastructure [2, 7].

Future frameworks must therefore evolve toward adaptive governance architectures—systems in which delegation authority, uncertainty interpretation, and ethical policy co-

evolve alongside the expanding frontier of materials knowledge.

Future research directions

Looking ahead, future research in decision authority frameworks for autonomous materials discovery systems should prioritize enhancements that address current limitations while leveraging emerging technologies in computational materials engineering [1, 2, 12]. A key direction involves advancing adaptive delegation models that incorporate real-time learning from human overrides, enabling AI to evolve its authority scope dynamically [5, 8, 16]. For example, integrating meta-learning with graph neural networks could allow systems to adjust delegations based on past intervention patterns, improving robustness in inverse design tasks [2, 13, 17].

Enhancing policy encoding through formal verification methods represents another promising avenue. By embedding policies as verifiable constraints in optimization loops, researchers can ensure compliance with evolving regulations, particularly in multimodal integration where data provenance is critical [12-14]. A conceptual formula for policy-augmented active learning might be:

$$a^* = argmax_a \left[I\left(\theta; y_a \mid D\right) - \mu C\left(a\right) \right] \quad (2)$$

where $I\left(\theta; y_a \mid D\right)$ is the mutual information for model update, C(a) quantifies constraint violations, and μ is a tunable penalty, delegating selection while enforcing policies [5, 8, 11]. This could facilitate safer autonomy in high-throughput computations [1].

Human-AI collaboration paradigms warrant exploration, such as gamified interfaces for overrides, fostering intuitive responsibility sharing [5, 20]. In autonomous laboratories, research into multi-agent orchestration could optimize delegations across distributed systems, as in international platforms [6, 11, 12, 16]. Addressing risks, future work should develop safeguarding mechanisms that prioritize human oversight in high-stakes scenarios, drawing from AI ethics literature [7, 9].

Interdisciplinary integration

Interdisciplinary efforts could merge materials informatics with cognitive science to model human-like decision authority, enhancing creativity in discovery loops [9, 20].

For nanomedicine and nanoparticle synthesis, directions include policy encodings for biocompatibility, enabling ethical delegations [3, 19].

Standardization and community

Initiatives Standardization of minimal working examples and community-driven benchmarks will accelerate progress, ensuring frameworks are testable across ecosystems [1, 10, 23]. Exploring large language models for automation in microscopy or phase mapping offers novel delegation opportunities [14, 22].

These directions promise to elevate autonomous systems toward reliable, ethical materials innovation [1, 2, 7].

Conclusion

In summary, decision authority frameworks are indispensable for the maturation of autonomous materials discovery systems within computational and data-driven materials engineering. By synthesizing delegation models, human override rights, responsibility assignment, and policy encoding, this review has illuminated their role in fostering efficient, safe workflows across materials informatics, closed-loop discovery, and inverse design. Despite challenges in uncertainty management, scalability, and ethics, the integrative perspectives drawn from recent literature underscore the frameworks' potential to bridge human expertise with AI autonomy. Future advancements in adaptive policies and interdisciplinary integrations will further solidify these systems as cornerstones of materials innovation, ultimately accelerating the development of sustainable, high-performance materials. This positions the field for transformative impacts, ensuring trustworthy autonomy in the pursuit of scientific breakthroughs.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 09 Mar 2025 Revised: 06 Apr 2025 Accepted: 16 Jun 2025
Published online: 18 September 2025

References

Stach E, DeCost B, Kusne AG, Hattrick-Simpers J, Brown KA, Reyes KG, et al. Autonomous experimentation systems for materials development: A community perspective. Matter. 2021;4(9):2702-26. https://doi.org/10.1016/j.matt.2021.06.036.

Tom G, Schmid SP, Baird SG, Cao Y, Darvish K, Hao H, et al. Self-Driving laboratories for chemistry and materials science. Chem Rev. 2024;124(16):9633-732. https://doi.org/10.1021/acs.chemrev.4c00055.

Abolhasani M, Kumacheva E. Self-driving laboratories: A paradigm shift in nanomedicine development. Matter. 2023;6(3):655-76.

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

Ahmadi M, Hao H, Aspuru-Guzik A. The future of self-driving laboratories: from human in the loop interactive AI to gamification. Digit Discov. 2024;3(4):621-36.

Pablo-García S, Skreta M, Lo S, Sim M, Rajaonson EM, Yoshikawa N, et al. ChemOS 2.0: An orchestration architecture for chemical self-driving laboratories. Matter. 2024;7(9):2959-77.

Qu X, Greenbaum D, Tang X, Jin Q, Zhu K, Yuan T, et al. Risks of AI scientists: Prioritizing safeguarding over autonomy. Nat Commun. 2025;16(8317). https://doi.org/10.1038/s41467-025-63913-1.

Adams F, McDannald A, Takeuchi I, Kusne AG. Human-in-the-loop for Bayesian autonomous materials phase mapping. Matter. 2024;7(3):988-97.

Ozin G. Homo deus: I am not a robot AI materials chemist. Matter. 2023;6(5):1324-6. https://doi.org/10.1016/j.matt.2023.03.035.

Hung L, Yager JA, Potocek 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(6):1273-9. https://doi.org/10.1039/D4DD00059E.

Kusne AG, McBride A. Scalable multi-agent lab framework for lab optimization. Matter. 2023;6(6):1880-93. https://doi.org/10.1016/j.matt.2023.03.022.

Pizzi G, Vogler M, Bieker J, Jørgensen PB, Häse F, Scherbela N, et al. Brokering between tenants for an international materials acceleration platform. Matter. 2023;6(9):2647-65. https://doi.org/10.1016/j.matt.2023.07.016.

Leong SX, Pablo-García S, Wong B, Aspuru-Guzik Á. MERMaid: Universal multimodal mining of chemical reactions from PDFs using vision-language models. Matter. 2025;8(12):102331.

Mandal I, Soni J, Zaki M, Gosvami NN, Krishnan NMA, Smedskjaer MM, et al. Evaluating large language model agents for automation of atomic force microscopy. Nat Commun. 2025;16(1):9104.

Karpovich C, Pan E, Jensen Z, Neilson J, Tamerler C, Aizenberg J, et al. An object-oriented framework to enable workflow evolution across materials acceleration platforms. Matter. 2022;5(11):3761-77.

Crebolder C, Aspuru-Guzik A, Pablo-García S, Skreta M, Yoshikawa N, Rajaonson EM, et al. El Agente: An autonomous agent for quantum chemistry. Matter. 2025;8(10):102193.

Salley D, Chakraborty R, Tsapatsis M, Bhan A, Josephson TR. Autonomous discovery of functional random heteropolymer blends through evolutionary formulation optimization. Matter. 2025;8(9):102289.

Savino E, Slattery A, Noël T. The role of flow chemistry in self-driving labs. Matter. 2025;8(7):102205.

Kim J, Park J, Han B, Jeong Y, Lee S, Kim H. Closed-loop optimization of nanoparticle synthesis enabled by robotics and machine learning. Matter. 2023;6(3):677-90.

Ozin GA. The curiosity-creativity element in HI-AI materials discovery. Matter. 2024;7(1):1-3.

Karpovich C, Pan E, Jensen Z, Neilson J, Tamerler C, Aizenberg J, et al. Challenges and opportunities for AI in synthetic solid-state inorganic chemistry. Matter. 2023;6(12):4304-26.

Na J, Kim S-J, Kim H, Kang S-H, Lee S. A unified microstructure segmentation approach via human-in-the-loop machine learning. Acta Mater. 2023;255:119038.

Baird SG, Sparks TD. What is a minimal working example for a self-driving laboratory? Matter. 2022;5(12):4170-8.