ORIGINAL RESEARCH **Open access**

A Conceptual Typology of Scientific Surprise in AI-Guided Discovery

Claire Martin[1], Julien Robert[2*], Sophie Bernard[1], Antoine Girard[2]

Abstract

The ambiguous usage of the term “surprise” in AI-guided discovery literature represents a significant conceptual barrier in artificial intelligence for materials science. Surprise is variously treated as a statistical anomaly flagged by machine learning models, a human psychological state of unexpectedness that prompts belief revision, an information-theoretic measure of divergence between prior and posterior beliefs, or an unexpected breakthrough that leads to a genuine scientific advance. This lack of precision confuses researchers, fragments the literature, and impedes the systematic design of AI systems capable of deliberately cultivating the forms of unexpectedness that drive materials innovation. This paper proposes a precise typology of scientific surprise consisting of four distinct types—predictive surprise, representational surprise, discovery surprise, and methodological surprise—tailored specifically to the domain of AI-guided discovery in materials science. The key distinctions among these types are articulated along four core dimensions: the source of the surprise (originating in the AI model or in the human scientist), the trigger (prediction error, out-of-distribution data, contradiction with existing theory, or unexpected patterns in the inquiry process itself), the experiencer (primarily the model or the scientist), and the epistemic consequences that follow (model retraining, expansion of representational capacity, theory revision, or redesign of search and measurement strategies). By furnishing this conceptual framework, the paper offers clear implications for designing AI systems that can report, distinguish, and cultivate productive forms of surprise, thereby transforming AI from a passive predictor into an active partner in the discovery process and enabling more effective, targeted responses to different kinds of unexpectedness in materials science.

Keywords Materials science, Scientific surprise, Conceptual typology, AI-guided discovery, Predictive surprise, Representational surprise

*Correspondence:
Julien Robert
julien.robert@gmail.com
[1] Department of Materials Informatics, University of Lyon, Lyon, France
[2] Department of AI Materials Systems, University of Strasbourg, Strasbourg, France

Introduction

“Surprise” appears throughout AI-guided discovery literature but means different things: statistical anomaly, human psychological state, information-theoretic measure, unexpected breakthrough. This conceptual ambiguity impedes systematic design for surprise cultivation [1-4].

The integration of artificial intelligence into materials science has transformed discovery workflows, enabling models to predict properties, suggest novel compositions, and optimize experimental campaigns at unprecedented scale [5, 6]. Butler *et al.* [5] demonstrate how machine learning can map vast chemical spaces and accelerate the identification of functional materials, while Schmidt *et al.* [6] review its successful application to solid-state systems ranging from batteries to photovoltaics. Zunger [7] further emphasizes that inverse design strategies succeed precisely when they uncover materials whose properties diverge from conventional expectations. Yet the very concept that often marks the transition from routine computation to genuine insight—surprise—remains ill-defined.

Cole *et al.* [1] long ago observed that surprise is epistemically significant because it forces revision of prior commitments; however, contemporary AI literature deploys the term in at least three incompatible registers. Some authors equate surprise with statistical outliers or novelty scores produced by anomaly-detection algorithms [4, 8-12]. Others treat it as a psychological or serendipitous event experienced by the human researcher [8-10]. Still others adopt the information-theoretic formulation of Bayesian surprise, quantifying it as the Kullback-Leibler divergence between prior and posterior distributions [2, 3]. These usages coexist without explicit differentiation, producing a literature in which the same word simultaneously denotes a model-internal signal, a human cognitive response, and a driver of scientific progress.

Such ambiguity is not merely semantic. When surprise is reduced to a numerical anomaly score, AI systems risk flagging countless statistically deviant but scientifically uninteresting data points, as illustrated in multiple materials-informatics studies that rely on unsupervised detection pipelines [13-17]. Conversely, when surprise is conceptualized exclusively as a human psychological state, the model's capacity to generate or suppress candidate surprises is overlooked, limiting the engineer's ability to design systems that actively promote epistemic gain [18-26]. In materials science, where the search space is combinatorially enormous and experimental validation is costly, the absence of a shared, precise vocabulary prevents researchers from asking targeted questions: which kind of surprise should the AI be optimized to produce, and which response protocol should follow?

The present paper, therefore, undertakes a boundary/definitional analysis. It first surveys the three dominant usages of "surprise" in the recent literature, then diagnoses the practical and theoretical problems that arise from their conflation, and finally introduces a four-type typology organized around two orthogonal dimensions—source of surprise (model versus scientist) and primary epistemic consequence (update to predictive machinery versus update to scientific understanding). By distinguishing predictive, representational, discovery, and methodological surprise, the framework supplies the field with a conceptual language that can guide the deliberate cultivation of productive surprise rather than its accidental occurrence. The remainder of the paper develops each type, examines its interrelations, and derives concrete implications for the architecture of next-generation AI discovery systems.

Surprise in Existing Literature

The notion of surprise within AI-guided materials discovery does not operate as a single, unified construct but instead traverses multiple conceptual domains that are rarely brought into explicit alignment. One influential lineage draws on information-theoretic and Bayesian principles, where surprise is formalized as the degree to which new observations revise prior beliefs. In this formulation, introduced by Mazzaglia *et al.* [2] and extended in subsequent work on attention mechanisms [3], surprise is quantified through the divergence between prior and posterior distributions, typically instantiated via the Kullback–Leibler metric. Although these contributions originate outside the immediate context of materials science, their logic persists in contemporary practice whenever belief updating is treated as a scalar signal within predictive models. Discussions of anomaly-driven discovery often invoke this intuition, as in the argument that deviations from expectation can catalyze scientific insight [4]. Yet, the mathematical grounding of such claims remains tethered—often implicitly—to this Bayesian framework.

A related but distinct interpretation emerges in the empirical literature, where surprise is operationalized through the detection of deviations from learned statistical regularities. Here, the emphasis shifts from belief revision to distributional inconsistency, and surprise becomes synonymous with anomaly or novelty relative to the model's training data. This interpretation has gained particular traction in materials informatics, where deep learning architectures and unsupervised techniques are deployed to identify out-of-distribution patterns. For instance, autoencoder-based approaches have been used to flag unusual diffraction signatures in large-scale X-ray datasets [12]. At the same time, analogous strategies extend to structural anomaly detection in fibrous systems [19], deviations in cable-force monitoring [16], and damage identification in fatigue analysis [18]. Similar methodologies appear across applications in nondestructive testing sensor networks [22], photovoltaic system diagnostics [23], and electron microscopy image degradation [21], consistently equating surprise with statistical rarity. The underlying assumption is that low-probability observations are inherently informative. Yet, this equivalence between deviation and significance often remains unexamined, raising questions about whether all anomalies carry meaningful scientific content.

Beyond these computational interpretations, a further dimension becomes visible when attention turns to the human experience of discovery, where surprise is intertwined with cognition, interpretation, and serendipity. Philosophical analyses emphasize that surprise does not arise solely from unexpected data but from the interaction between chance events and a prepared interpretive framework [8, 9]. This perspective is echoed in studies of human–machine collaboration, where systems are designed not only to detect anomalies but to amplify opportunities for insight by human observers [10]. Empirical work in domains such as astronomy illustrates how machine learning can surface unusual patterns, yet the recognition of their significance ultimately depends on human interpretation [25]. Parallel discussions in information science further highlight the role of serendipitous encounters in shaping knowledge acquisition, whether in social-media-driven drug repurposing [26] or in broader models of information-seeking behavior [27]. The extension of these ideas into system design, as proposed by Reviglio [28], reframes surprise as an experiential and architectural property rather than a purely computational output.

What emerges from the coexistence of these perspectives is a conceptual fragmentation that remains largely unaddressed within the literature. It is not uncommon for a single study to move fluidly between formal metrics of belief updating [2, 3], operational definitions rooted in anomaly detection [12, 17], and narratives of human insight or discovery [4, 25], without clarifying how these distinct meanings of surprise relate to one another. This fluidity, while often productive at the level of application, obscures the mechanisms through which surprise functions at different stages of the discovery process. The result is a body of work that is empirically rich yet theoretically diffuse, underscoring the need for a more systematic articulation of how these interpretations intersect, diverge, and can be coherently integrated within AI-driven materials science.

The Problem with Current Usage

The coexistence of divergent meanings of “surprise” within AI-guided materials discovery introduces a set of conceptual tensions that constrain both theoretical precision and system design. A particularly consequential issue arises from the implicit conflation of statistical deviation with epistemic significance. When surprise is defined solely in terms of low probability under a model’s current parameterization, it is tacitly assumed that such deviation corresponds to meaningful scientific insight. Yet this alignment is far from guaranteed. A model may encounter an observation that lies at the tail of its learned distribution while remaining fully consistent with established physical theory, offering little in the way of conceptual advancement [4]. Conversely, a result may appear entirely routine from the model’s perspective while destabilizing entrenched assumptions within the scientific community, thereby generating genuine epistemic disruption, as emphasized in classical accounts of belief revision [1]. The failure to distinguish between these cases leads to a misalignment of objectives, where AI systems are optimized to amplify internal measures of surprise rather than to cultivate outcomes that advance scientific understanding.

This misalignment is further compounded by the marginalization of the human interpretive role within many computational frameworks. When surprise is reduced to a scalar anomaly score, the epistemic labor of the scientist—interpreting, contextualizing, and assigning significance—is effectively abstracted away. Such abstraction obscures the fact that surprise is not an intrinsic property of data alone but emerges through the interaction between observation and prior knowledge. Studies of serendipity underscore that identical observations can elicit radically different responses depending on the observer’s conceptual readiness and disciplinary background [8, 9]. Within collaborative human–machine systems, this variability is not incidental but constitutive of discovery itself [10, 27]. A pipeline that indiscriminately flags deviations without regard to their interpretive potential risks saturates the user with signals that fail to translate into meaningful insight. Over time, this can erode trust in the system and divert experimental effort toward phenomena that are statistically unusual yet epistemically trivial.

A related limitation lies in the tacit assumption that greater surprise is inherently desirable. This presumption overlooks the heterogeneous nature of deviations encountered in practice. Many anomalies arise not from novel physical phenomena but from noise, measurement error, or artifacts introduced by incomplete or biased training data [11, 13, 19]. Without a conceptual framework capable of distinguishing generative deviations from spurious ones, system designers lack the means to encode preferences that privilege scientifically productive outcomes. The absence of such differentiation complicates the specification of loss functions, acquisition strategies, and

retraining protocols, as there is no principled basis for determining whether a detected anomaly should trigger model adaptation, theoretical reconsideration, or simple rejection as noise. In effect, the system becomes responsive to deviation without discrimination, blurring the boundary between discovery and artifact.

Proposed Typology of Scientific Surprise

This ambiguity is reinforced by the lack of shared terminology across studies, which impedes cumulative progress. Terms such as “surprise,” “anomaly,” and “outlier” are often used interchangeably, despite referring to phenomena that differ in both origin and consequence. As a result, insights generated within one methodological context are difficult to transfer or generalize to others, limiting the development of coherent design principles. The cost of this fragmentation is particularly acute in materials science, where data acquisition is resource-intensive, and the identification of genuinely novel phenomena carries substantial scientific and practical value [5, 6, 20]. Under these conditions, the absence of conceptual clarity is not merely an academic concern but a barrier to effective system engineering.

Figure 1 presents the proposed typology as a branching decision structure linking the source of surprise and its primary epistemic consequence to four distinct forms of scientific surprise.

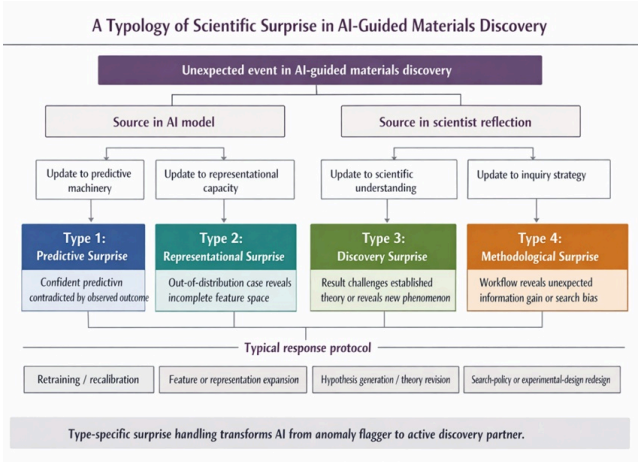

Figure 1. The proposed typology is a branching decision structure linking the source of surprise and its primary epistemic consequence to four distinct forms of scientific surprise.

The typology can be conceptualized as a comparison matrix described here in full sentences. The rows represent the source of surprise: the first row corresponds to surprises that originate within the AI model’s internal computations; the second row corresponds to surprises experienced primarily by the human scientist reflecting on model outputs or experimental results. The columns represent primary epistemic consequence: the first column denotes consequences that principally update the model’s ability to predict future observations; the second column denotes consequences that update the scientist’s theoretical understanding or the strategic design of the discovery process itself.

In the top-left cell (model source, predictive update) sits type 1 Predictive Surprise: the model registers a large divergence between its confident prediction and the observed outcome, prompting parameter adjustment or retraining. In the top-right cell (model source, understanding update) sits type 2 Representational Surprise: the model encounters data that falls outside its learned distribution, revealing the incompleteness of its current feature space or embedding and necessitating expansion of representational capacity. In the bottom-right cell (scientist source, understanding update) sits type 3 discovery surprise: the scientist encounters a result that contradicts established theory or reveals a previously unknown phenomenon, triggering hypothesis generation and potential paradigm shift. In the bottom-left cell (scientist source, predictive-strategy update) sits type 4 methodological surprise: the scientist realizes that the very strategy of inquiry—active-learning policy, experimental design, or measurement protocol—has produced an unexpected pattern of information gain, prompting revision of how future searches are conducted.

Along the four required dimensions, the types differ sharply.

Table 1 differentiates the four surprise types across trigger, detection locus, epistemic consequence, and response protocol, clarifying why each requires a distinct intervention.

Table 1. Analytical differentiation of the four types of scientific surprise in AI-guided materials discovery

Dimension	Type 1: predictive surprise	Type 2: representational surprise	Type discov surpri

Primary source	AI model	AI model	Scient
Immediate trigger	High-confidence prediction contradicted by observation	Out-of-distribution case or latent-space failure	Resu contrad establis theory expectat
Locus of detection	Residual error, uncertainty calibration, and loss spike	Density collapse, reconstruction error, and embedding distance	Huma interpret of the significa of the re
What is being challenged	Parameter fits within the existing model	Adequacy of feature space or representation	Doma theor assumpt or classifica schem
Primary epistemic consequence	Model retraining or recalibration	Expansion or refactoring of the representation	Hypothe generati theor revisi
Typical response protocol	Retrain, recalibrate, and validate the measurement	Add descriptors, expand ontology, and retrain architecture	Paus automat interpr theorize, reframe mechan
Main risk if misclassified	Overfitting to noise	Treating ontology failure as ordinary error	Premat retraini that suppres genuir discov
Value for AI-system design	Improves predictive fidelity	Improves domain coverage and robustness	Improv scienti creativity explana advanc

Source distinguishes model-internal signals (types 1 and 2) from scientist-centered reflection (types 3 and 4). Trigger for type 1 is a low-probability outcome under high model confidence; for type 2, it is the absence of analogous training instances; for type 3, it is incompatibility with domain theory; for type 4, it is an unanticipated information-yield profile of the search policy itself. Experiencer is the model for types 1 and 2 (quantified via confidence or density estimates) and the scientist for types 3 and 4 (subjective recognition of novelty or strategic misalignment). Epistemic consequence is correspondingly local (model update for type 1), representational (embedding expansion for type 2), theoretical (belief revision for type 3), or strategic (inquiry redesign for type 4).

This matrix resolves the literature's conflations by making explicit which cell each previous usage occupies. Bayesian formulations [2, 3] map most naturally onto type 1, anomaly-detection pipelines [12, 17] straddle types 1 and 2, while serendipity accounts [8, 25] emphasize type 3. Methodological surprise (type 4) has been largely invisible, yet it is crucial for closing the loop between model behavior and experimental strategy. The typology, therefore, supplies the missing vocabulary, enabling designers to specify which type of surprise an AI system should cultivate and which response protocol should follow. Subsequent sections elaborate on each type individually before examining their dynamic interrelations.

Predictive Surprise

Type 1 Predictive Surprise occurs when a model's confident prediction diverges sharply from the observed outcome, registering high prediction error on a point the model considered well-constrained. The trigger is the realization of a low-probability event under the model's current posterior; the experiencer is first the model itself (via internal uncertainty or loss signals) and secondarily the scientist who observes the mismatch. The epistemic consequence is targeted model update or retraining to restore predictive fidelity.

This type aligns closely with the Bayesian surprise tradition [2, 3] yet is distinguished by its focus on the model's own confidence rather than raw information gain. In materials applications, a model might predict a bandgap of 2.1 eV for a candidate compound with narrow uncertainty; experimental measurement returns 3.4 eV, forcing gradient updates that refine the learned mapping between structure and property [5, 6]. The surprise is predictive because the model's internal belief state is directly contradicted, yet it

need not imply any deeper representational inadequacy or theoretical rupture.

Importantly, predictive surprise can be both useful and misleading. When systematic, it signals regions where the training distribution is insufficiently dense, prompting beneficial retraining [12]. When sporadic, it may reflect experimental noise or measurement error, yielding no lasting epistemic gain [4]. The typology, therefore, insists that designers distinguish predictive surprise from the other three types rather than treating every large residual as automatically valuable. A system that merely maximizes predictive surprise risks overfitting to outliers instead of expanding scientific understanding.

In practice, predictive surprise is the easiest type to quantify and therefore the most commonly engineered. Confidence-aware architectures already report prediction uncertainty; augmenting them with an explicit "predictive-surprise flag" would allow the human user to decide whether to trigger immediate retraining, defer action pending further checks, or route the event to another surprise-handling module. Such differentiation prevents the conflation that currently hampers the literature and prepares the ground for the more complex forms of surprise examined in subsequent sections.

Representational Surprise

Type 2 Representational Surprise arises when a new data point falls outside the model's learned training distribution, revealing that the current representational framework is fundamentally incomplete rather than merely miscalibrated on a specific prediction. The trigger is the absence of analogous local environments, compositions, or measurement regimes in the training corpus, so the model cannot assign a meaningful density or embedding to the incoming observation. This form of surprise is experienced jointly by the model, which registers low confidence or high reconstruction error in its internal latent space, and by the scientist, who recognizes the novelty as a signal that existing descriptors or featurizations fail to capture essential aspects of the material. The primary epistemic consequence is the imperative to expand or refactor the model's representational capacity, for instance, by incorporating new graph neural network layers, augmenting the feature set with higher-order structural invariants, or retraining on an enlarged chemical space.

In materials informatics, this type is frequently encountered when models trained exclusively on binary compounds confront ternary or quaternary systems whose local coordination environments have no precedent [12]. Banko *et al.* [12] document precisely such cases in large X-ray diffraction datasets, where deep autoencoders flag patterns whose embeddings lie far from the learned manifold, prompting the realization that the original featurization omitted critical multi-body interactions. Similar dynamics appear in anomaly detection pipelines applied to fibrous media [19] and electron-microscope images of rubber composites [21], where the surprise is not a simple prediction error but an indication that the entire embedding space must be enlarged to accommodate previously unseen microstructural motifs.

Representational surprise is conceptually distinct from predictive surprise because it does not presuppose that the model ever made a confident forecast; instead, the model withholds confidence altogether, signaling a boundary condition of its knowledge representation. Where type 1 prompts parameter tuning within a fixed architecture, Type 2 demands architectural or ontological revision—redefining what counts as a "feature" in the materials domain. This distinction matters deeply for AI system design: an algorithm optimized solely for minimizing predictive residuals may remain silent in the face of true representational gaps, whereas a system engineered to monitor embedding density can actively surface these gaps before they propagate into downstream discovery failures [17, 22].

The scientist's role further differentiates this type. Upon encountering an out-of-distribution sample, the researcher must decide whether to treat the event as noise or as an invitation to enrich the model's ontology, perhaps by synthesizing targeted validation compounds or by integrating domain knowledge from related subfields such as high-entropy alloys or two-dimensional heterostructures [20]. Without explicit labeling of the surprise as representational, however, the literature risks conflating it with mere predictive mismatch, leading to suboptimal responses such as aggressive retraining on noisy outliers rather than deliberate expansion of the representation space [11, 13]. By isolating representational surprise as a distinct category, the typology equips designers with a clear criterion: when density estimates collapse, the appropriate intervention is not incremental gradient descent but a higher-order update to the model's conceptual vocabulary for materials structure. This reframing transforms out-of-

distribution detection from a defensive filtering mechanism into a proactive engine for representational growth, directly addressing one of the central bottlenecks in scaling AI-guided discovery across the vast compositional space of modern materials science.

Discovery Surprise

Type 3 Discovery Surprise is realized when an experimental or computational outcome contradicts established scientific beliefs or discloses a previously unknown phenomenon, irrespective of the model's internal probability estimates. The trigger is incompatibility between the observed result and the prevailing theoretical or empirical consensus within the materials community, not merely deviation from a learned statistical distribution. This surprise is experienced primarily by the scientist, whose background knowledge and expectations are directly challenged. At the same time, the model may register only low confidence or even remain agnostic if the result lies within its broad uncertainty bounds. The epistemic consequence is theory revision, hypothesis generation, or the opening of an entirely new research trajectory, moving beyond model calibration toward genuine advancement of domain understanding.

Philosophically, this type resonates with Levi's classical account of surprise as the force that compels revision of prior commitments [1]. In the AI-materials context, discovery surprise manifests when a predicted stable crystal structure fails to appear experimentally, and an unanticipated polymorph emerges instead, or when a material exhibits an electronic property that defies band-theory expectations despite accurate structural prediction [4, 25]. Giles and Walkowicz [25] illustrate the phenomenon in astronomical anomaly pipelines where the model flags outliers that later prove to be new classes of objects; the true surprise resides not in the model's score but in the human recognition that existing classification schemes are inadequate. Analogous episodes occur in serendipity-driven materials work, where citizen-science platforms or human–machine collaborations surface results that reorient entire subfields [10, 26].

Crucially, discovery surprise can occur without any preceding predictive or representational surprise from the AI system. A model may correctly forecast a property value. Yet, the scientist experiences profound surprise because the outcome falsifies a long-standing assumption about structure–property relationships across an entire chemical family [8, 9]. This independence underscores why equating all surprise with model-internal signals is limiting: the model may be statistically unsurprised while the scientific community undergoes a genuine epistemic rupture. Conversely, great predictive surprise that remains unexplained can evolve into discovery surprise once the scientist supplies the missing theoretical link [27].

The typology, therefore, positions discovery surprise as the culminating epistemic payoff rather than an intermediate modeling artifact. It demands that AI systems be designed not merely to minimize surprise for the model but to surface candidate discoveries in ways that maximize the probability of human theoretical insight. By labeling an event explicitly as a discovery surprise, the framework guides researchers to shift from immediate model retraining to reflective activities such as literature re-examination, analogy construction across domains, or formulation of new mechanistic hypotheses. In this manner, the typology restores the scientist to the center of the discovery loop, ensuring that AI serves as a catalyst for human creativity rather than a substitute for it.

Methodological Surprise

Methodological surprise occurs when the discovery process itself (e.g., active learning or optimization) produces unexpected information gain, revealing flaws or opportunities in the search strategy. It is experienced through reflection on the workflow rather than individual predictions. For instance, Bayesian optimization may repeatedly select "unpromising" regions that yield high value [5], exposing better search paths [5, 28, 29]. Unlike representational surprise, which expands features, this type reshapes the decision process guiding experiments [6, 20]. Its response is to recalibrate or redesign the optimization policy rather than retrain the model.

Relationship between Types

The four surprise types are interconnected. Predictive surprise can escalate into discovery surprise when errors reveal deeper theoretical failure [1, 4]. Representational surprise can trigger methodological surprise by exposing sampling bias [12, 17]. Discovery surprise feeds back into predictive surprise when a new theory generates new testable predictions. These transitions form iterative loops,

making the typology a guide for diagnosing and responding to unexpected outcomes.

Table 2 shows that surprise types are dynamically linked through identifiable transition pathways, allowing designers to specify escalation rules rather than treating unexpectedness as isolated events.

Table 2. Transition pathways, escalation risks, and design interventions across surprise types

Transition pathway	Mechanism of movement	Diagnostic signal	Epistemic if ignor
Predictive → Discovery	Repeated model failures reveal deeper theoretical inconsistency	Persistent residual pattern across related compounds or conditions	Treatin paradig relevar anomalie mere calib error
Representational → Methodological	Out-of-distribution events reveal biased sampling or constrained search policy	Novel cases cluster in neglected regions of materials space	Repeate expandi features w correcti workflow
Discovery → Predictive	New theoretical insight is encoded into the model and generates new testable predictions	Emergence of a revised mechanism, descriptor, or hypothesis	Failure operation concept advance future pred cycles
Methodological → Representational	Workflow redesign exposes previously unobserved regions requiring new descriptors	New sampling regime reveals latent structures absent from current feature space	Improve search w unchang ontology y interpret bottlene
Within-type repetition: Predictive	Residual errors accumulate without conceptual escalation	Same class of mismatch recurs after retraining	Endless l tuning with epistemic
Within-type repetition: Representational	Recurrent OOD detections indicate systematic ontology limits	Multiple embedding failures in adjacent domains	Fragmer patches in of coher representa redesig
Within-type repetition: Discovery	Similar theoretical tensions recur across cases	Multiple findings strain the same background assumption	Misse opportunit synthesi paradig articulat
Within-type repetition: Methodological	Search-policy anomalies repeatedly recur	Persistent optimizer bias or asymmetric information yield	Chroni inefficien the autono discove pipelin

Implications for Ai-Guided Discovery

AI systems should classify surprise by type and apply targeted responses: retraining for predictive [2, 3], feature expansion for representational [12, 21], human intervention for discovery [8, 25], and policy optimization for methodological [5]. This avoids treating all surprises uniformly and improves efficiency. The framework also clarifies roles: models handle predictive/representational surprises, while humans guide discovery/methodological ones [9, 27]. Overall, it enables deliberate cultivation of productive surprise in AI-driven science.

Conclusion

This paper has proposed a precise typology of scientific surprise consisting of four distinct types—predictive,

representational, discovery, and methodological—organized around the orthogonal dimensions of source and epistemic consequence. The typology clarifies that Bayesian formulations map primarily onto predictive surprise, anomaly-detection pipelines straddle predictive and representational surprise, serendipity scholarship centers discovery surprise, and methodological surprise has until now remained largely unarticulated yet essential for closing the discovery loop.

The ultimate implication is a shift from AI that merely minimizes predictive error toward systems deliberately designed to cultivate the full spectrum of productive surprise, thereby accelerating the pace and depth of materials innovation. Future work should focus on operationalizing the typology in autonomous laboratories and validating its utility through reflective case studies of past discoveries re-analyzed through this four-type lens. In doing so, the community can move beyond vague invocations of “surprise” toward a mature science of engineered serendipity in which artificial intelligence and human creativity reinforce each other at every stage of the discovery process.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 02 Dec 2021 Revised: 15 Feb 2022 Accepted: 15 Mar 2022
Published online: 18 July 2022

References

Cole SR, Edwards JK, Greenland S. Surprise! Am J Epidemiol. 2021;190(2):191-3.

Mazzaglia P, Catal O, Verbelen T, Dhoedt B. Curiosity-driven exploration via latent Bayesian surprise. Proc AAAI Conf Artif Intell. 2022;36(7):7752-60.

Chieppe P, Sweetser P, Newman E. Bayesian modelling of the well-made surprise. In: ICCC; 2022. p. 126-35.

Marchese E, Caldarelli G, Squartini T. Detecting mesoscale structures by surprise. Commun Phys. 2022;5(1):132.

Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A. Machine learning for molecular and materials science. Nature. 2018;559(7715):547-55.

Schmidt J, Marques MR, Botti S, Marques MA. Recent advances and applications of machine learning in solid-state materials science. npj Comput Mater. 2019;5(1):83.

Zunger A. Inverse design in search of materials with target functionalities. Nat Rev Chem. 2018;2(4):0121.

Copeland S. On serendipity in science: Discovery at the intersection of chance and wisdom. Synthese. 2019;196(6):2385-406.

Ippoliti E. Scientific discovery reloaded. Topoi. 2020;39(4):847-56.

Trouille L, Lintott CJ, Fortson LF. Citizen science frontiers: Efficiency, engagement, and serendipitous discovery with human-machine systems. Proc Natl Acad Sci U S A. 2019;116(6):1902-9.

Cha YJ, Wang Z. Unsupervised novelty detection-based structural damage localization using a density peaks-based fast clustering algorithm. Struct Health Monit. 2018;17(2):313-24.

Banko L, Maffettone PM, Naujoks D, Olds D, Ludwig A. Deep learning for visualization and novelty detection in large X-ray diffraction datasets. npj Comput Mater. 2021;7(1):104.

Rossi A, Montefoschi F, Rizzo A, Diligenti M, Festucci C. Auto-associative recurrent neural networks and long term dependencies in novelty detection for audio surveillance applications. In: IOP Conference Series: Materials Science and Engineering; 2017;261(1):012009.

Sun Y, Yan G, Shi X. Anomaly detection algorithm based on electric equipment. In: IOP Conference Series: Materials Science and Engineering; 2019;631(4):042046.

Lehr J, Philipps J, Hoang VN, Wrangel DV, Krüger J. Supervised learning vs. unsupervised learning: A comparison for optical inspection applications in quality control. In: IOP Conference Series: Materials Science and Engineering; 2021;1140(1):012049.

Li S, Niu J, Li Z. Novelty detection of cable-stayed bridges based on cable force correlation exploration using spatiotemporal graph convolutional networks. Struct Health Monit. 2021;20(4):2216-28.

Bao Y, Tang Z, Li H, Zhang Y. Computer vision and deep learning-based data anomaly detection method for structural health monitoring. Struct Health Monit. 2019;18(2):401-21.

Kalia S, Zeitler J, Mohan CK, Weiss V. Machine learning and anomaly detection algorithms for damage characterization from compliance data in three-point bending fatigue. J Nondestruct Eval Diagn Progn Eng Syst. 2021;4(4):041011.

Dresvyanskiy D, Karaseva T, Mitrofanov S, Redenbach C, Schwaar S, Makogin V, et al. Application of clustering methods to anomaly detection in fibrous media. In: IOP Conference Series: Materials Science and Engineering; 2019;537(2):022001.

Frydrych K, Karimi K, Pecelerowicz M, Alvarez R, Dominguez-Gutiérrez FJ, Rovaris F, et al. Materials informatics for mechanical deformation: A review of applications and challenges. Materials. 2021;14(19):5764.

Togo R, Saito N, Ogawa T, Haseyama M. Estimating regions of deterioration in electron microscope images of rubber materials via a transfer learning-based anomaly detection model. IEEE Access. 2019;7:162395-404.

Kraljevski I, Duckhorn F, Tschöpe C, Wolff M. Machine learning for anomaly assessment in sensor networks for NDT in aerospace. IEEE Sens J. 2021;21(9):11000-8.

Xie X, Wei X, Wang X, Guo X, Li J, Cheng Z. Photovoltaic panel anomaly detection system based on unmanned aerial vehicle platform. In: IOP Conference Series: Materials Science and Engineering; 2020;768(7):07206.

Tanuska P, Spendla L, Kebisek M, Duris R, Stremy M. Smart anomaly detection and prediction for assembly process maintenance in compliance with Industry 4.0. Sensors. 2021;21(7):2376.

Giles D, Walkowicz L. Systematic serendipity: A test of unsupervised machine learning as a method for anomaly detection. Mon Not R Astron Soc. 2019;484(1):834-49.

Ru B, Li D, Hu Y, Yao L. Serendipity-A machine-learning application for mining serendipitous drug usage from social media. IEEE Trans Nanobioscience. 2019;18(3):324-34.

Liu Y, Qin C, Ma X, Liang H. Serendipity in human information behavior: A systematic review. J Doc. 2022;78(2):435-62.

Reviglio U. Serendipity as an emerging design principle of the infosphere: Challenges and opportunities. Ethics Inf Technol. 2019;21(2):151-66.

Oliynyk AO, Buriak JM. Virtual issue on machine-learning discoveries in materials science. Chem Mater. 2019;31(20):8243-7.