ORIGINAL RESEARCH **Open access**

Human-in-the-Loop without the Hype: A Conceptual Taxonomy of Human Roles in Materials AI

Nguyen Thanh Huy[1*], Le Thi Bich[2], Pham Quang Minh[1]

Abstract

Human-in-the-loop (HITL) approaches are increasingly invoked in materials artificial intelligence (AI) as a presumed remedy for unreliable models, opaque predictions, and domain-shift failures. Yet “including a human” often functions as a rhetorical assurance rather than a precise scientific claim, masking the fact that humans participate in materially different ways: as labelers, judges, curators, constraint designers, hypothesis framers, risk owners, and accountability anchors. This conceptual manuscript argues that HITL is not a single method but a family of epistemic and governance roles that shape what an AI output means, what it can justify, and what actions it can responsibly warrant. Building on recent developments in materials informatics, active learning, uncertainty quantification, interpretable machine learning, and scientific machine learning, we synthesize a theory-first view of human involvement as a structured intervention in the AI-to-decision pathway rather than an informal override mechanism. We introduce a novel taxonomy that distinguishes (i) where humans intervene in the pipeline (data, representation, model, evaluation, decision), (ii) what kind of authority they exert (epistemic, normative, operational), and (iii) how their involvement changes the legitimacy of downstream claims under differing stakes. The resulting framework replaces HITL hype with a falsifiable conceptual vocabulary for designing responsibility, reliability, and restraint in materials AI.

Keywords Materials informatics, Interpretability, Decision-making, Trustworthy AI, Uncertainty, Human-in-the-loop

*Correspondence:
Nguyen Thanh Huy
huy.nguyen@gmail.com
[1] Department of Materials Science and Data Engineering, Faculty of Engineering, Vietnam National University, Hanoi, Vietnam
[2] Department of Artificial Intelligence Systems, Faculty of Engineering, Can Tho University, Can Tho, Vietnam

Introduction

Artificial intelligence has become an enabling language for contemporary materials science: it compresses high-dimensional structure–property relationships into predictive surrogates, proposes candidates in vast compositional spaces, and accelerates screening across heterogeneous datasets. The materials AI literature increasingly presents these systems as design engines rather than merely regression tools, shifting their role from descriptive modeling to prescriptive decision support. As this shift occurs, the cost of error becomes contextual rather than purely statistical: a model can be “accurate on average” yet unsafe for rare but catastrophic failure modes in high-stakes deployment contexts (e.g., structural alloys, battery safety, or critical process windows). This mismatch between predictive adequacy and decision legitimacy has made “human-in-the-loop” (HITL) an almost universal recommendation for trustworthy materials AI [1-4].

Yet the current discourse around HITL is often conceptually imprecise. In many papers, a human is introduced as if their presence alone suffices to repair epistemic fragility: a domain expert “checks” outputs, confirms plausibility, or selects candidates. Such descriptions conflate at least three different ideas: (i) human oversight as governance, (ii) human

correction as error reduction, and (iii) human participation as knowledge production. In practice, these are distinct contributions. A human can improve dataset fidelity without improving mechanistic meaning; can improve interpretability without improving reliability; or can override decisions for safety even when the model is highly reliable. Without a structured taxonomy, “human-in-the-loop” becomes an elastic label that promises rigor while evading specification of responsibility, authority, and scope.

This paper proposes a deliberately non-hyped reframing: HITL is best understood as a role system rather than a technique. The central claim is that human involvement is not merely a supplement to computation but a structured intervention in the pipeline of epistemic warrant—the conceptual pathway by which model outputs become actionable knowledge. In materials AI, the model’s output is rarely the final scientific object; it is an intermediate artifact interpreted under domain assumptions, constraints, and decision stakes. Humans, therefore, do not “add intuition” abstractly; they determine which assumptions are valid, which constraints are binding, and which risks are acceptable. HITL is a governance architecture for interpretation and action, not a generic trust injection.

This is not a call to “bring back” human expertise as an alternative to machine learning. Materials science has always been human-in-the-loop: experimental design, microstructure interpretation, phase diagram reasoning, and property trade-off selection are inherently human-guided tasks. What has changed is that AI systems now operate at a scale and abstraction that can detach decision-making from direct empirical contact. In high-throughput, data-centric workflows, model outputs can circulate as if they were facts rather than proposals, leading to inflated claims and premature operationalization. When AI predictions are deployed into synthesis planning, candidate ranking, or process optimization, the human role is no longer incidental; it is the primary mechanism by which the system maintains epistemic humility and decision discipline [2, 5].

However, HITL can also fail in predictable ways. One failure mode is responsibility laundering, where human presence is treated as a liability shield (“an expert approved it”) while the system’s epistemic limits remain unexamined. Another is cognitive overload, where humans are asked to validate thousands of candidates without structured support, effectively replacing automation with fatigue. A third is semantic overreach, where model explanations or saliency maps are treated as mechanistic evidence, and the human becomes an interpreter of artifacts that were never designed to support mechanistic inference. Finally, authority mismatch occurs when humans are asked to decide matters outside their mandate—such as accepting a prediction that requires probabilistic calibration expertise, or rejecting an output that is correct but counterintuitive.

These failure modes reveal a deeper issue: HITL is not inherently beneficial. It is beneficial only when the human role is well-defined, properly scoped, and correctly placed relative to the AI pipeline’s failure points. For example, a human who corrects training labels exerts influence upstream at the level of data truth. A human who sets safety thresholds exerts influence downstream at the level of decision policy. A human who selects descriptors or invariances influences representation—and therefore controls what the model is even capable of learning. These are categorically different “loops,” and they should not be treated as interchangeable.

This manuscript addresses this conceptual gap by introducing a taxonomy of human roles in materials AI that is designed to be (i) materials-specific, (ii) decision-aware, and (iii) epistemically explicit. The taxonomy differentiates human intervention by location (data, representation, model, evaluation, decision), by authority type (epistemic, normative, operational), and by stake sensitivity (the burden of justification required before action). In doing so, we convert HITL from a slogan into a design language: a framework for assigning responsibility, auditing epistemic warrant, and preventing overreach.

Importantly, this paper is purely conceptual. It does not propose new algorithms, human-interface designs, or experimental protocols. Instead, it offers a theory-first roadmap that clarifies what HITL means in the context of materials AI, and how to reason about its legitimacy claims. The novelty lies not in listing common HITL practices, but in developing

a coherent model of how human roles restructure the AI-to-action chain under uncertainty, interpretability limitations, and domain shift.

Theoretical Background & Literature Synthesis

Why human roles became a central question in materials AI

Materials AI has matured from predicting individual properties to coordinating multi-stage workflows that traverse data acquisition, representation, modeling, and decision selection. In these workflows, model outputs function as decision triggers: selecting compounds for synthesis, prioritizing experiments, or ranking microstructures for targeted performance. This amplifies the importance of epistemic robustness because errors propagate nonlinearly: a small prediction bias can reshape search trajectories and concentrate effort in misleading regions of design space [1, 6-8].

At the same time, material datasets are heterogeneous in terms of provenance and meaning. Labels can encode different operational definitions (e.g., “strength” measured under varied strain rates), and data can mix theoretical and experimental sources with incompatible assumptions. Such heterogeneity is not merely noise—it is a semantic mismatch. Humans are often required to interpret whether two records are commensurate, whether a surrogate can generalize across processing regimes, or whether a descriptor set respects known symmetries. These are conceptual judgments about validity, not simple data cleaning tasks [6, 9].

This helps explain why HITL is repeatedly recommended: the human is presumed to reintroduce meaning, context, and scientific discipline. Yet without a defined role taxonomy, humans are treated as universal solvents: plug-in fixes for uncertainty, bias, and interpretability. The literature shows that these concerns are distinct and require different forms of intervention.

Human-in-the-loop vs. human-on-the-loop vs. human-in-command

A first conceptual clarification is that “human involvement” has multiple governance positions. In AI safety and accountability discourse, a common separation is:

· Human-in-the-loop: the system requires human input during operation (e.g., labeling, selection, approval).

· Human-on-the-loop: the system runs autonomously but is monitored and can be overridden.

· Human-in-command: humans define the objectives and acceptable constraints, even if execution is automated.

In materials AI, these distinctions matter because AI systems are often embedded in scientific workflows rather than consumer-facing products. A model may not be “operating” continuously; it may be generating candidate lists, suggesting measurement targets, or proposing design rules. Human-in-command roles (objective setting, constraint definition, risk framing) can dominate the workflow’s legitimacy even when humans are not approving frequently. Conversely, even with frequent human approvals, objectives can still fail if they are ill-posed (e.g., optimizing a proxy metric that diverges from the true property of interest) [2, 10].

Thus, a materials-centered HITL theory must not only ask whether humans are involved, but also how they govern meaning: by shaping the target, validating the label, selecting the hypothesis class, or enforcing decision thresholds.

Uncertainty is not just a model output; It is a decision boundary

Uncertainty quantification (UQ) has become a cornerstone of trustworthy machine learning discourse and is increasingly adopted in materials informatics [5, 7]. However, UQ is often interpreted narrowly as a numerical interval around a

prediction. A role-based view suggests a deeper claim: uncertainty is a boundary object that connects model knowledge to human responsibility.

Materials science exhibits multiple uncertainty sources:

· Aleatory uncertainty: variability intrinsic to processes (e.g., microstructural heterogeneity).

· Epistemic uncertainty: model ignorance due to limited data coverage or incomplete physics.

· Semantic uncertainty: ambiguity in labels, targets, or the meaning of “success” under varying conditions.

Human roles differ depending on which uncertainty dominates. When epistemic uncertainty dominates, the key human role may be designing acquisition strategies, deciding what information would reduce ignorance, or recognizing that extrapolation is occurring. When semantic uncertainty dominates, the human role shifts to defining the concept precisely: what does “high corrosion resistance” mean across pH, temperature, and microstructure? This is not a model calibration issue but a scientific definition issue. Therefore, HITL cannot be framed purely as reducing uncertainty; it must be framed as interpreting which uncertainty matters for action [3, 11].

Interpretability is not a substitute for explanation

Interpretability methods are frequently proposed to make models more trustworthy by exposing the “reasons” behind predictions [12, 13]. Yet interpretability does not automatically produce mechanistic understanding. Many interpretability methods reveal correlations inside a learned representation; they do not establish causal structure. In materials contexts, this distinction is high-stakes: a feature attribution map may highlight a compositional descriptor associated with performance, but that does not mean it is a mechanism or a stable design rule across processing conditions.

This is precisely where human roles become subtle. Humans can misinterpret interpretability outputs as mechanistic evidence, and then propagate a false narrative into design or publication. Alternatively, humans can use interpretability properly as a diagnostic—checking for spurious drivers (dataset artifacts), confirming alignment with known invariances, or identifying representation fragility. Thus, the human role in interpretability is not “reading explanations,” but auditing the epistemic status of explanations: are these signals predictive, causal, or merely correlational? [2, 14]

A rigorous HITL taxonomy must therefore separate:

· The human as meaning-maker (translating artifacts into scientific language), and

· The human as validator (ensuring that translation is epistemically justified).

Active learning and Bayesian optimization: where human choice becomes scientific control

Active learning and Bayesian optimization are widely adopted paradigms in materials discovery because they aim to allocate measurement effort efficiently under limited budgets [4, 8]. These paradigms often appear “automated,” but they embed human decisions at multiple layers: choice of objective, constraints, acquisition policy, and acceptance criteria. Humans often control what counts as improvement (e.g., property maxima, trade-offs, feasibility) and what risks are tolerable (e.g., avoiding toxic chemistries, excluding unstable phases).

From a conceptual viewpoint, this means the human is not merely a label provider. Humans are policy designers: they govern how uncertainty is translated into action (exploration vs. exploitation) and how multi-objective trade-offs are resolved. This policy-level role can dominate downstream outcomes more than model architecture choices. Yet it is rarely explicitly recognized as a human role and is instead treated as “problem setup.” In a HITL taxonomy, “problem setup” is

reclassified as a formal role: the human as objective author and constraint legislator, which is neither purely epistemic nor purely operational, but normative and scientific at once [10, 15].

Scientific machine learning and the boundary between physics and judgment

Physics-informed and scientific machine learning approaches aim to incorporate known constraints, symmetries, or governing equations to improve generalization and reduce data needs [16, 17]. In materials AI, this includes enforcing invariances, embedding physical priors, or constraining plausible microstructural transformations. These approaches elevate a different human role: the human as prior architect. The researcher decides which physical constraints are valid, how strongly they should bind the model, and which simplifications are acceptable.

This role is fundamentally epistemic: it encodes what the community considers legitimate physics knowledge. But it is also fallible: an incorrect prior can create overconfident failure. Therefore, a HITL framework must treat “adding physics” not as an objective upgrade, but as a human-authored intervention that must be audited for scope and validity. In other words, the human can increase trustworthiness—or manufacture confident error—through prior choice [2, 17].

Trustworthiness as a system property: Humans as accountability anchors

Trustworthy AI is increasingly framed as a system property rather than a model metric: trust depends on transparency, robustness, uncertainty awareness, and governance [11, 18]. In materials AI, trustworthiness must also incorporate scientific accountability: who is responsible for a candidate selection that leads to wasted synthesis cycles, or for a publication claim that overstates generality?

A role-based account locates the human not only as an enhancer of model performance, but as an accountability anchor: the agent who signs off on claims, carries reputational responsibility, and enforces restraint. This is especially important for “silent failures” in materials AI: cases where the model produces plausible outputs that are scientifically unjustified due to domain shift, label mismatch, or missing processing variables. In these cases, human restraint—declining to claim or to act—is an epistemic virtue, not a system weakness [19, 20].

Summary: Why we need a taxonomy rather than another workflow diagram

The literature demonstrates a consistent pattern: human roles are invoked to address uncertainty, interpretability, generalization, and governance. But these are not the same problem, and they require distinct interventions. What is missing is a unified conceptual language that classifies human participation as a structured set of roles with specific authority types and stakeholder sensitivity. Without such a taxonomy, HITL becomes a slogan that obscures rather than clarifies responsibility. The next section proposes such a framework.

Proposed conceptual framework

Overview: HITL as a “Role–Authority–Stake” system

We propose a novel conceptual framework, the Role–Authority–Stake (RAS) Taxonomy, which defines human involvement in materials AI not as an informal layer of “oversight,” but as a structured system governed by three orthogonal axes: role location, authority type, and stake sensitivity. First, the RAS taxonomy specifies *where* humans intervene in the pipeline by identifying five primary control locations: Data, Representation, Model, Evaluation, and Decision. At the data stage, human intervention includes collection choices, labeling decisions, cleaning, curation, and documentation practices that determine whether the dataset is sufficiently coherent to support later inference. At the representation stage, humans shape the choice of descriptors, invariance assumptions, and feature validity, directly influencing what the model is even capable of expressing. At the modeling stage, humans intervene through architecture selection, prior assumptions, objective formulation, and constraint definitions, thereby deciding what the system is trained to value and penalize. At evaluation, humans define validation logic, interpret uncertainty, and stress-test failure modes to determine whether performance claims can legitimately transfer beyond training conditions. Finally, at the decision stage,

humans set deployment thresholds, accept or reject trade-offs, and assign accountability, thereby translating model outputs into actions that may carry real-world risk.

Second, the RAS taxonomy distinguishes the kinds of control humans exercise into three authority types. Epistemic authority concerns judgments about what is valid, such as whether a label represents a stable material concept or whether a model output can support a given claim. Normative authority concerns what *should* be optimized or avoided, such as binding constraints related to safety, sustainability, feasibility, or ethical limits. Operational authority concerns what will happen next in the workflow, including which candidates to pursue, what experiments to run, and whether a system output is approved for downstream use. Rather than treating "human involvement" as a single category, this authority axis explains why adding a human reviewer does not automatically increase trustworthiness unless the human is empowered within the correct authority domain.

Third, the RAS taxonomy treats human involvement as inseparable from stake sensitivity, recognizing that the same intervention can be legitimate under low-stakes conditions and illegitimate under high-stakes conditions. Under exploratory stakes, speculative candidate generation is acceptable because the purpose is ideation rather than commitment. Under translational stakes, outputs may be used only conditionally and must be paired with targeted checks and validation logic before they affect real decisions. Under critical stakes, only highly warranted actions are permissible, and silence (abstention) is not a failure but a correct scientific outcome when applicability is uncertain. Taken together, these three axes produce the central conceptual benefit of the RAS design: it prevents the category error of treating HITL as a unitary guarantee of trust, and forces a sharper question instead—which role, with which authority, under which stakes, is being invoked as the basis for action?

The taxonomy: Seven human roles without overlap

Within this three-axis system, the framework defines seven conceptually non-redundant human roles that are frequently conflated in HITL discourse. Role 1, the curator, operates primarily at the data location and controls dataset inclusion criteria, resolves conflicts across sources, and enforces semantic consistency so that the training base is not internally incoherent. Role 2, the label legitimizer, exercises epistemic authority by judging whether labels correspond to stable material concepts, including whether measurements are commensurable across instruments, protocols, and reporting conventions. Role 3, the representation steward, governs meaning through representation choices by selecting descriptors and invariances that determine what the model can represent and what it must ignore, thereby preventing representation fragility and semantic distortion. Role 4, the constraint legislator, expresses normative authority by defining feasibility, safety, sustainability, and domain restrictions as binding decision boundaries rather than optional preferences. Role 5, the uncertainty interpreter, functions at the evaluation–decision boundary and treats uncertainty not as a number but as a permission structure, deciding whether outputs warrant deployment, conditional use, or abstention. Role 6, the claim auditor, combines epistemic and communicative authority by separating predictive signals from mechanistic narratives and preventing interpretability artifacts from being mistaken for explanation or causal insight. Role 7, the risk owner, carries accountability authority by defining acceptable loss, failure tolerance, stopping conditions, and responsibility assignment when decisions are made under stakes.

A key claim of the framework is that these roles are not reducible to generic "domain expertise," because expertise does not guarantee correct authority placement or valid role execution. A senior experimentalist may be well-positioned as a risk owner due to experience with failure consequences, yet may not be trained to interpret uncertainty as a boundary condition rather than a confidence score. Conversely, an ML specialist may competently interpret uncertainty behavior yet lack legitimate normative authority to encode safety or sustainability constraints that are institutionally and ethically governed. The RAS taxonomy, therefore, reframes HITL as a role-based architecture rather than an informal appeal to expertise.

What the framework explains: When HITL helps and when it is "confidence theater"

The RAS taxonomy explains HITL success through a precise causal logic: HITL improves materials AI outcomes when role–failure matching is explicit, when the authority type used to resolve the failure is legitimate, and when stakes are clearly specified rather than assumed. Under this framework, HITL helps when the correct role intervenes at the correct pipeline location—for example, when a semantic mismatch in labels is corrected by a label legitimizer rather than misdiagnosed as a model interpretability problem. HITL also helps when authority alignment is enforced, meaning that normative constraints must be authored as constraints and policies rather than retrospectively inferred from model behavior. Finally, HITL helps only when stakeholder conditions are explicit, because critical decisions require stronger warrants than exploratory ranking does, and the threshold for action must rise with consequence.

At the same time, the framework explains systematic "HITL failure by design," including failure modes that persist even when humans are present. Confidence theater occurs when humans provide superficial approval without epistemic audit, functioning as rhetorical legitimacy rather than scientific control. Role overload occurs when one individual is expected to fulfill multiple roles simultaneously—curation, legitimacy, constraint policy, uncertainty interpretation, and risk ownership—creating bottlenecks and weakening the epistemic quality of decisions. Authority mismatch occurs when a human is asked to justify what the system cannot warrant, such as treating interpretability overlays as mechanistic evidence or expecting intuition to replace abstention logic under domain shift. Silent failure denial occurs when abstention is treated as unacceptable even when scientific legitimacy requires silence, forcing the system into overreach rather than restraint. Through these explanations, the RAS taxonomy reclassifies HITL not as a patch for weak generalization, but as a controlled architecture for responsibility, legitimacy, and action under bounded warrant **(Figure 1).**

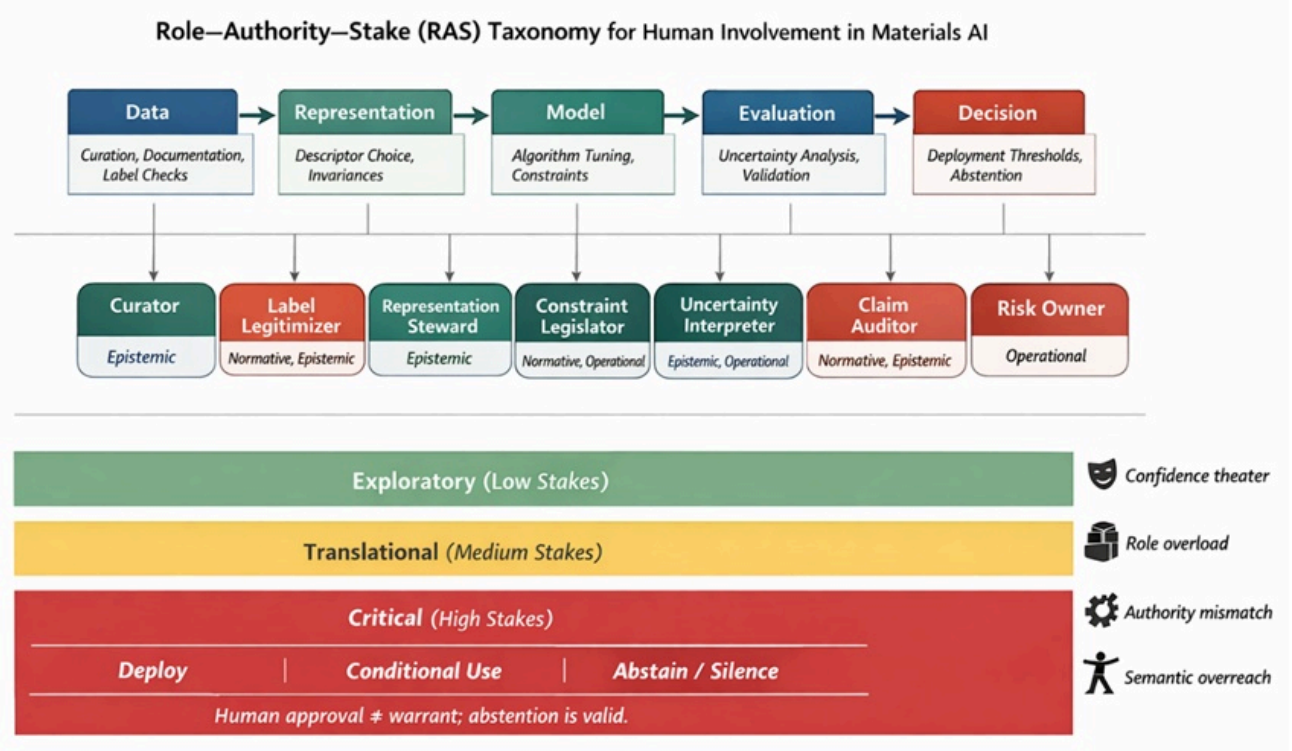

Figure 1. Role–authority–stake (RAS) taxonomy for human involvement in materials AI"

Propositions

This section formalizes the RAS conceptual framework into falsifiable, manuscript-grade propositions. Importantly, these claims are not implementation prescriptions and do not assume particular algorithms, datasets, or experimental regimes. Instead, they operate at the theoretical level, specifying how human involvement alters the epistemic status of model outputs and the legitimacy of decisions based on them. In doing so, the propositions reframe HITL away from a

performance optimization trope and toward a responsibility architecture for materials AI—one that distinguishes between what is predicted, what is warranted, and what can be legitimately acted upon.

Proposition 1

Claim

Human-in-the-loop (HITL) in materials AI cannot be meaningfully defined as a single technique; it is best understood as a family of role-specific interventions whose epistemic consequences differ across pipeline locations, authority types, and stake levels.

Rationale

The term HITL is routinely used as a blanket descriptor for fundamentally non-equivalent interventions: dataset curation, labeling legitimacy checks, representation design, constraint definition, uncertainty interpretation, claim auditing, and decision approval. When these heterogeneous activities are compressed into a single category, “HITL” becomes conceptually inflated into a proxy for responsibility itself, enabling the rhetorical move that “we had a human in the loop” is sufficient evidence of scientific legitimacy. This proposition follows governance-oriented HITL scholarship, which stresses that interaction structure and placement matter more than vague calls for oversight [21, 22]. In materials informatics, the need for disaggregation is sharper because labels, targets, and evaluation contexts can be semantically unstable across processing histories and measurement conditions, meaning that “human involvement” without role clarity can conceal rather than resolve epistemic fragility [1, 6].

Proposition 2

Claim

The primary scientific value of HITL in materials AI is not improved predictive performance, but modified epistemic warrant: the legitimacy of treating an output as a basis for action or claim.

Rationale

In decision-centric scientific workflows, accuracy alone is an incomplete criterion for legitimacy. A prediction can be numerically correct on benchmark evaluations yet remain illegitimate to deploy if its validity conditions are unknown, its applicability domain is unclear, or the failure costs are asymmetric and high. The RAS framework emphasizes that certain human roles—particularly the risk owner and constraint legislator—can change what the system is allowed to do under uncertainty without altering the model’s internal mapping. This implies that HITL should be evaluated in terms of warrant transitions, such as moving from signal to candidate, from candidate to conditional recommendation, and from recommendation to actionable decision, rather than being reduced to regression improvements [11, 18]. This proposition is consistent with the broader view that trustworthiness is a system property shaped by uncertainty handling, decision framing, and governance placement, not merely by the choice of model family [5, 7, 11].

Proposition 3

Claim

In materials AI, HITL failures are more strongly predicted by authority mismatch (the wrong human role exercising the wrong kind of authority) than by model architecture choice.

Rationale

Common explanations for HITL failure assign blame to the algorithmic substrate, often framed as “black-box deep learning” versus “transparent modeling.” The RAS taxonomy predicts a different failure generator: structural misassignment. A domain expert may be asked to validate probabilistic calibration that they cannot audit, a data scientist may be expected to legislate sustainability constraints without legitimate normative authority, or a reviewer may be pressured to treat feature attributions as mechanistic evidence. These breakdowns are not fundamentally algorithmic; they are failures of governance placement and authority type. HITL surveys repeatedly show that human-in-the-loop

outcomes depend on interaction design and control placement rather than on generic “human oversight” [21, 22]. In materials AI, the consequences of authority mismatch are amplified because interpretability artifacts can be mistaken for explanation, creating a pathway from prediction to unjustified mechanism claims—an epistemically dangerous escalation that the RAS framework treats as claim inflation [12, 13].

Proposition 4

Claim

The strongest—and least explicitly acknowledged—reason HITL is repeatedly demanded in materials AI is semantic uncertainty, not merely epistemic uncertainty.

Rationale

Semantic uncertainty arises when targets and labels carry ambiguous or shifting meanings across datasets, processing histories, measurement protocols, and reporting conventions. Materials science is unusually vulnerable to such drift because many evaluation terms are operationally defined rather than ontologically fixed. Concepts such as “stability,” “durability,” “high performance,” and even “strength” often encode implicit experimental conditions, time scales, or failure definitions. Under these conditions, a model can be epistemically confident while semantically misaligned, producing outputs that appear valid but do not map to the intended material concept. Humans are uniquely positioned to resolve semantic mismatch through definition work, concept alignment, and scope policing—interventions that cannot be substituted by post hoc uncertainty intervals alone [6, 9]. This proposition clarifies why “more data” frequently fails: the limiting factor is not scarcity but conceptual inconsistency, and scaling incoherence produces confusion at scale.

Proposition 5

Claim

Interpretability tools increase the risk of scientific overreach unless paired with a claim-auditor role that distinguishes predictive correlation from mechanistic explanation.

Rationale

Interpretability methods can create an aesthetic of scientific meaning by producing human-readable attributions, saliency maps, or feature rankings. In materials AI, this readability can be mistaken for mechanism, encouraging mechanistic storytelling even when the model’s internal logic is correlation-bound and non-causal. Without an explicit claim-auditor role, interpretability becomes a narrative generator that increases the probability of semantic overreach: the output is treated not merely as predictive guidance but as explanation. This is not only a philosophical concern; it influences what is written, believed, funded, and deployed. Interpretability scholarship has repeatedly warned that explanations can be misleading, incomplete, or misused when treated as ground truth [12, 13, 23]. The RAS taxonomy, therefore, treats interpretability as safe only when embedded within a governance structure that includes claim auditing, rather than positioned as an automatic legitimacy amplifier.

Proposition 6

Claim

A mature HITL system must treat abstention (“model silence”) as a legitimate human-authorized outcome, not a failure of the pipeline.

Rationale

When a model operates outside its applicability domain, the appropriate response is often not “try anyway with a human check,” but refusal to act or formal downgrading of claim strength. This is especially true in high-stakes material contexts where invalid extrapolation can lead to costly, unsafe, or irrecoverable consequences. The RAS framework clarifies that human authority is not only the authority to approve decisions; it is also the authority to block action and impose epistemic restraint. A HITL pipeline that lacks the capacity to justify silence is structurally biased toward overreach, treating outputs

as needing to be “used” rather than “warranted” [11, 19, 20]. In this sense, abstention is not the absence of intelligence but the expression of scientific legitimacy in the face of uncertainty.

Proposition 7

Claim

HITL requirements should scale with decision stakes: the same loop can be justified in low-stakes exploration but unjustified in high-stakes deployment unless strengthened.

Rationale

In exploratory screening workflows, humans may legitimately accept uncertain rankings and heuristic signals as idea generators, since the goal is not commitment but hypothesis expansion. In translational settings, similar uncertainty levels should trigger conditional use paired with targeted validation rather than open deployment. In critical deployment contexts, the same uncertainty levels must trigger abstention or formal verification gates, because responsibility cannot be reduced to optimism when the consequences of failure are significant. HITL, therefore, cannot be specified independently of stake context, and universal recommendations to “always include a human in the loop” are scientifically weaker than stake-indexed governance designs. This proposition aligns with contemporary perspectives that emphasize uncertainty-aware decision-making and calibrated trust, rather than generic trust as a global good [5, 7, 11]. By treating stakes as a first-order variable, the RAS taxonomy replaces HITL hype with decision-legitimate HITL design.

Results and Discussion

Why “human-in-the-loop” became hype

“HITL” gained popularity because it appears to reconcile two competing pressures: the demand for accelerated discovery and the fear of unreliable AI. Its rhetoric suggests a simple compromise: automation with safety. But the slogan obscures the fact that “the human” is not a uniform entity, and “the loop” is not a single location. In materials AI, these ambiguities generate a false sense of rigor: adding a human does not necessarily increase scientific validity; it may instead increase the appearance of accountability.

From the perspective developed here, HITL is hype when it serves as a trust label rather than a role specification. The RAS framework converts HITL into a falsifiable claim: which human role, exerting which authority, at which stake level, changes which warrant boundary? Without this specificity, HITL descriptions are structurally non-auditable.

What the taxonomy changes in materials AI practice (conceptually)

Although this manuscript is intentionally non-implementation-focused, a conceptual framework still has operational consequences: it changes what people think they are doing when they build “human-centered” pipelines.

(i) From “expert check” to “warrant engineering.”

The central conceptual upgrade is to treat human involvement as engineering the warrant pathway rather than checking outputs. A label legitimizer ensures semantic stability. A constraint legislator prevents optimization from violating physical or ethical feasibility. A claim auditor constrains scientific language. A risk owner defines acceptable loss and stopping logic. These are different forms of scientific work that deserve explicit recognition.

(ii) From “trust” to “permission structures.”

Trust is psychologically ambiguous; permission is structurally explicit. The pipeline outcome is not “trust the model,” but one of: deploy, conditionally use, or abstain. This aligns with uncertainty-centered approaches to decision-making [5, 7].

(iii) From “human intuition” to “human accountability.”

The RAS framework shifts emphasis away from intuition and toward accountability: the system must have a risk owner even if the model is accurate. Accountability is not a property of the model; it is a property of the sociotechnical system [11, 18].

Four failure modes of HITL in materials AI (and why they recur)

Failure mode A — Confidence theater

Humans provide approval signals that create an illusion of validation without epistemic substance. For example, an expert skims a ranked list and endorses it because it "looks plausible," while uncertainty and applicability conditions remain undefined. This is theater because it converts human authority into a stamp, not a warrant analysis.

Failure mode B — Semantic overreach

Humans translate predictive correlations into mechanistic claims because interpretability outputs offer tempting narratives. When a model highlights a feature correlated with performance, humans overinterpret it as causal. This is particularly common in materials contexts where domain language invites mechanism talk. Interpretability research cautions that explanations can be misleading without careful framing [12, 13, 23-27].

Failure mode C — Role overload and bottlenecking

A single "domain expert" is expected to label, curate, interpret uncertainty, validate mechanisms, and decide actions. This collapses distinct roles and undermines each of them. In real research groups, role overload creates fragile systems: human review becomes inconsistent, delayed, and susceptible to confirmation bias.

Failure mode D — Authority mismatch

Humans are placed into roles they cannot legitimately perform. Example: a computational scientist is asked to define operational safety constraints for industrial deployment, or an experimentalist is asked to judge statistical calibration. Authority mismatch is dangerous because it produces confident decisions with weak epistemic grounding.

When HITL genuinely helps: A "role–failure match" logic

The RAS framework suggests that human-in-the-loop (HITL) succeeds only when it is designed as an explicit role–failure match, rather than treated as a generic "expert oversight" layer. In this view, the value of human involvement depends on whether the *right* human role is placed at the right point of the pipeline to correct a *specific* failure mode. For instance, when the dominant failure is label inconsistency, the appropriate remedy is the Label Legitimizer, because the core problem is epistemic legitimacy in the target definition—not model transparency or interpretability. Similarly, when the failure is a domain shift, the remedy is not "expert intuition," but a structured pairing of uncertainty interpretation with an explicit abstention policy, because the system must be able to recognize when it lacks warrant rather than compensate with informal confidence. When the failure is optimization misalignment, the correct intervention is the Constraint Legislator, since the underlying issue is normative and operational mis-specification of what the system is allowed to optimize—not a shortage of data. Finally, when the failure is claim inflation, the remedy is the Claim Auditor, because scientific overreach is not corrected by higher predictive fit (e.g., higher R^2), but by restricting which claims are warranted given the available evidence and boundary conditions. This logic aligns with HITL scholarship, which emphasizes that trustworthiness arises from the interaction structure and the placement of control, not from the mere presence of a human somewhere in the system [21, 22]. It also reinforces a system-level view of trustworthiness in which governance is distributed across roles and decision points rather than collapsed into an undefined "human check" [11, 18, 28-30].

Stake sensitivity: The missing variable in HITL debates

Many HITL debates implicitly treat human involvement as categorically beneficial, as if more human presence automatically improves responsibility. The framework developed here rejects that assumption and instead treats HITL as

stake-indexed, meaning that the same human behavior can be valid in one context but epistemically invalid in another. In low-stakes exploratory screening, humans may legitimately accept heuristic rankings, weak signals, and fast interpretations because the system is functioning as an idea generator rather than a decision authority. In translational settings, however, humans must shift from heuristic acceptance to structured caution, requiring conditional checks, targeted validation, and explicit boundary-awareness before advancing claims or recommendations. In critical deployment contexts, the expected human role becomes even stricter: humans must actively enforce abstention whenever applicability is uncertain, because the cost of error is no longer informational but consequential. The implication is that HITL is not a guarantee of responsibility; rather, HITL is a mechanism for calibrating responsibility to stakes, and it only works when the human role, authority type, and decision posture are explicitly aligned with the consequences of being wrong [28-32].

Conceptual implications for autonomy and “self-driving laboratories”

Materials AI is increasingly discussed alongside autonomous discovery systems, closed-loop design, and “self-driving laboratories.” Although this manuscript does not address robotic automation, experimental execution, or hardware integration, the conceptual implication still holds: autonomy does not eliminate humans—it redistributes human authority upward. As systems become more autonomous, humans shift away from routine operational contributions, such as manual labeling, toward higher-level governance roles, including constraint setting, interpretive control, and responsibility assignment. In the RAS view, autonomy therefore amplifies the importance of the Constraint Legislator, because optimization processes can rapidly move into scientifically invalid or unsafe regions if constraints are absent or underspecified. It also elevates the role of the Uncertainty Interpreter, because automated loops can magnify invalid assumptions and propagate them through repeated iterations unless uncertainty is explicitly interpreted as a boundary condition for action. Finally, it increases the centrality of the Risk Owner, because accountability cannot be automated away; responsibility must remain identifiable, attributable, and decision-linked. Under this framing, HITL is not the opposite of autonomy. Instead, HITL functions as the governance skeleton that makes autonomy scientifically legitimate, ensuring that faster optimization does not become faster epistemic overreach [31-35].

Limitations of the present conceptual framework

This work deliberately excludes implementation-level design, such as interfaces, elicitation protocols, cognitive ergonomics, and organizational incentives. It also does not claim that any particular algorithmic approach is superior. The contribution is taxonomic and epistemic: a conceptual architecture for describing human roles without collapsing them into “expert review.”

A second limitation is that the framework does not provide empirical measures for role quality. However, the value of a conceptual taxonomy lies in enabling precise critique: it allows reviewers and readers to ask which role exists, where it is placed, what authority it carries, and how it supports the warranted claim.

Conclusion

Human-in-the-loop has become a default phrase in materials artificial intelligence, frequently invoked as a trust-enhancing ingredient without adequate conceptual specification. This manuscript argues that such usage produces more reassurance than rigor. In materials AI, the legitimacy of an output depends not only on predictive quality but on semantic alignment, uncertainty interpretation, applicability limits, and decision stakes. These properties are not repaired by “adding a human” in a generic sense.

To address this gap, we proposed the role–authority–stake (RAS) taxonomy, a theory-first framework that reframes HITL as a structured system of human roles—curator, label legitimizer, representation steward, constraint legislator, uncertainty interpreter, claim auditor, and risk owner. The framework distinguishes where humans intervene in the AI pipeline, what kind of authority they exercise, and how decision stakes transform what counts as responsible action. This structure

clarifies why HITL sometimes succeeds (role–failure matching) and why it often fails (confidence theater, semantic overreach, authority mismatch, role overload). Most importantly, it makes explicit that abstention and silence are legitimate outcomes in high-stakes contexts: a mature HITL design must permit refusal, not only approval.

The conceptual shift advocated here is from “trust the model because a human checked it” to “define permission structures that convert uncertainty into disciplined action.” By providing a falsifiable vocabulary for human roles, the RAS taxonomy enables sharper scientific communication, more accountable workflow design, and more principled restraint in the face of uncertainty. It also offers a foundation for future work that can empirically study role quality and interaction design, without treating HITL as a monolithic solution. In this sense, human-in-the-loop should not be marketed as a cure for AI limitations; it should be treated as a rigorous architecture for epistemic warrant and responsibility in accelerated materials discovery.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 14 Nov 2023 Revised: 30 Dec 2023 Accepted: 01 Feb 2024
Published online: 18 July 2024

References

Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, et al. The materials project: A materials genome approach to accelerating materials innovation. APL Mater. 2021;9(3):030902.

Draxl C, Scheffler M. The NOMAD laboratory: From data sharing to artificial intelligence. J Phys Mater. 2020;3(4):040302.

Chen C, Ye W, Zuo Y, Zheng C, Ong SP. Graph networks as a universal machine learning framework for molecules and crystals. Chem Mater. 2019;31(9):3564–72.

Merchant A, Batzner S, Schoenholz SS, Aykol M, Cheon G, Cubuk ED. Scaling deep learning for materials discovery. Nature. 2023;624:80–5.

Gawlik A, Balachandran PV. Machine learning in materials science: recent progress and emerging applications. MRS Bull. 2021;46:1–8.

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

Frazier PI. A tutorial on Bayesian optimization. https://arxiv.org/pdf/1807.02811. 2018.

Settles B. Active learning literature survey. Univ Wisconsin-Madison. 2010.

Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Inf Fusion. 2021;76:243–97.

Huang B, von Lilienfeld OA. Ab initio machine learning in chemical compound space. Chem Rev. 2021;121(16):10001–36.

Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1:206–15.

Molnar C. Interpretable Machine Learning. Leanpub; 2022.

Mosqueira-Rey E, Hernández-Pereira E, Alonso-Ríos D, Bobes-Bascarán J, Fernández-Leal Á. Human-in-the-loop machine learning: a state of the art. Artif Intell Rev. 2023;56:3005–54.

Wu X, Xiao L, Sun Y, Zhang J, Ma T, He L. A survey of human-in-the-loop for machine learning. Future Gener Comput Syst. 2022;135:364–81.

Isayev O, Oses C, Toher C, Gossett E, Curtarolo S, Tropsha A. Universal fragment descriptors for predicting properties of inorganic crystals. Nat Commun. 2020;11:2593.

Karniadakis GE, Kevrekidis IG, Lu L, Perdikaris P, Wang S, Yang L. Physics-informed machine learning. Nat Rev Phys. 2021;3:422–40.

Brunton SL, Kutz JN. Data-driven science and engineering: Machine learning, dynamical systems, and control. 2nd ed. Cambridge: Cambridge University Press; 2022.

Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv: Machine Learning; 2017. https://doi.org/10.48550/arXiv.1702.08608.

Ghorbani A, Wexler J, Zou J, Kim B. Towards automatic concept-based explanations. Proceedings of the 33rd International Conference on Neural Information Processing Systems; 2019. pp. 9277-86. https://doi.org/10.48550/arXiv.1902.03129.

Xu Q, Zhang H, et al. Data-driven materials discovery and design using machine learning. Patterns. 2021;2(11):100332.

Lin S, Huang W, et al. Human-in-the-loop learning: A review of methods and applications. ACM Comput Surv. 2021;54(8):1–36.

Amershi S, Weld D, Vorvoreanu M, Fourney A, Nushi B, Collisson P, et al. Guidelines for human-ai interaction. In: Proceedings of the 2019 chi conference on human factors in computing systems (chi '19). New York (NY): Association for Computing Machinery; 2019. Paper 3:1-13.
https://doi.org/10.1145/3290605.3300233.

Arya V, Bellamy RK, Chen PY, Dhurandhar A, Hind M, Hoffman SC, et al. One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. 2019.
https://doi.org/10.48550/arXiv.1909.03012.

Zhou Z, et al. Transfer learning in materials informatics: A review. Comput Mater Sci. 2020;186:110014.

Jablonka KM, Ongari D, Moosavi SM, Smit B. Big-data science in porous materials: materials genomics and machine learning. Chem Rev. 2020;120(16):8066–129.

Moosavi SM, Jablonka KM, Smit B. The role of machine learning in the understanding and design of materials. J Am Chem Soc. 2020;142(48):20273–87.

Ward L, Wolverton C. Atomistic calculations and materials informatics: A review. Curr Opin Solid State Mater Sci. 2020;24(3):100805.

Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Phys Rev Lett. 2020;120:145301.

Chen S, et al. A critical review of machine learning of defects in materials. Prog Mater Sci. 2022;127:100951.

Goodall REA, Lee AA. Predicting materials properties without crystal structure: deep representation learning from stoichiometry. Nat Commun. 2020;11:6280.

Tshitoyan V, Dagdelen J, Weston L, Dunn A, Rong Z, Kononova O, et al. Unsupervised word embeddings capture latent knowledge from materials science literature. Nature. 2019;571(7763):95-8.
https://doi.org/10.1038/s41586-019-1335-8.

Himanen L, Geurts A, Foster AS, Rinke P. Data-driven materials science: status, challenges, and perspectives. Adv Sci. 2020;7(21):1900808.

Wang AY-T, Murdock R, Kauwe S, Oliynyk AO, Gurlo A, Brgoch J, et al. Machine learning for materials scientists: An introductory guide toward best practices. Chem Mater. 2020;32(12):4954–65.

Yang Z, et al. A perspective on active learning in materials science. NPJ Comput Mater. 2021;7:1–11.

Lookman T, Balachandran PV, Xue D, Yuan R. Active learning in materials science with emphasis on adaptive sampling using uncertainties. NPJ Comput Mater. 2022;8:1–13.