Discover Artificial Intelligence

REVIEW Open Access

Interaction-induced knowledge narrowing risk in LLM systems

Dale Rutherford[1] and Ningning Wu[1*]

*Correspondence:
Ningning Wu
nxwu@ualr.edu
[1]Department of Information Science, University of Arkansas at Little Rock, 2801 S. University Ave, Little Rock, AR 72204, USA

Abstract

AI governance frameworks address well-documented risks through controls on data, models, and post-deployment performance. This paper identifies a distinct, under-governed lifecycle risk arising not from model error or degradation, but from interaction itself. Termed Interaction-Induced Knowledge Narrowing (IIKN), this risk occurs when bounded context windows, ranked probabilistic synthesis, and iterative interaction dynamics systematically constrain the epistemic scope surfaced during AI system use. IIKN manifests as unobserved exclusion rather than incorrectness: AI systems may produce fluent, accurate, and aligned outputs while silently omitting plausible, defensible, and consequential alternatives. Because IIKN leaves no error signal, it routinely passes undetected by conventional assurance mechanisms. This paper provides a formal definition, differentiates IIKN from adjacent risk categories, and explains its emergence through inference-time selection and interaction path dependence. Mapping IIKN across the AI lifecycle demonstrates why late-stage review is insufficient and early controls are necessary. The paper proposes proportionate governance controls, including epistemic scope declaration, breadth-before-depth interaction patterns, coverage confirmation checkpoints, and lifecycle-aligned trigger gates, that render exclusion explicit and accountable without requiring exhaustive enumeration. Potential alignment with ISO/IEC 42,001 and the NIST AI Risk Management Framework is demonstrated, with the argument that governing what AI systems omit is as critical as governing what they produce, particularly as organizations adopt increasingly agentic systems.

Keywords AI governance, Large language models, Epistemic risk, Human-AI interaction, Decision support systems, Lifecycle governance, Interaction design, Knowledge management

1 Introduction and problem framing

AI governance frameworks have advanced rapidly in response to well-established risks associated with large language model systems. Training data bias, hallucination, model drift, privacy exposure, and security vulnerabilities are now widely recognized, formally categorized, and addressed through regulatory guidance, internal risk registers, and enterprise control frameworks [12, 32, 7]. These risks share a common structural orientation: they are framed as defects in data, models, or post-deployment performance, and

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

they are mitigated through controls applied during training, validation, deployment, or monitoring.

This framing reflects an implicit assumption that once a model is trained, evaluated, and deployed within acceptable thresholds, the primary remaining governance concerns relate to correctness, alignment, and degradation over time [12, 25, 29]. What it leaves structurally unexamined is the epistemic impact of interaction itself: how inference-time mechanisms and human-AI interaction dynamics shape which information, options, and perspectives are surfaced, considered, and ultimately acted upon during system use.

Modern AI systems operate within bounded context windows and rely on ranked probabilistic synthesis to determine which responses to present to users [31, 26]. These mechanisms are commonly treated as technical implementation details or performance constraints. In practice, they function as active selection and exclusion mechanisms at inference time. Context window capacities vary by orders of magnitude across widely deployed systems, ranging from roughly two hundred thousand to a theoretical ten million tokens, as Table 1 illustrates with three representative tiers rather than an exhaustive or ranked comparison; these figures are point-in-time, change frequently across providers, and are cited to their primary sources per row [2, 22, 19]. Published technical documentation for major deployed systems confirms that attention mechanisms, retrieval filters, and summarization pipelines each introduce selection pressure prior to output generation, a structural property established in the transformer architecture literature [31], retrieval-augmented generation frameworks [17, 14], and model-card and internal-audit documentation practices adopted by major AI providers [20, 27]. The governance implication is not the exact size of a context window, which is why Table 1 is deliberately illustrative rather than comprehensive, but the degree to which inference-time selection pressure is distributed across attention, retrieval, ranking, summarization, and memory mechanisms. Larger context windows do not eliminate epistemic constraint; they shift selection pressure upstream into retrieval, ranking, and compression processes, consistent with broader critiques of what scaling does and does not resolve in large language systems [5], and with infrastructure-studies accounts of how capacity constraints relocate rather than disappear [28].

A targeted review of the AI governance literature conducted for this paper, encompassing ACM FAccT proceedings (2018–2025), Springer AI and Ethics, AAAI AI safety publications, and the principal regulatory frameworks including ISO/IEC 42,001 [15], the NIST AI RMF [21], and the EU AI Act [11], found no existing governance construct

Table 1 Illustrative comparison of context window capacity across architectural layers (three representative tiers; figures are point-in-time as of publication and cited to primary sources per row; the argument of this paper does not depend on precise figures or their rank order)

Illustrative tier	Model	Layer 1: Attention (Context window, source)	Layer 2: Conversation	Layer 3: Retrieval/RAG
Smaller, widely deployed	Claude Opus 4.5	200,000 tokens [2]	Tool-based retrieval; no auto-inject	Projects auto-RAG beyond 200 K; retrieves 5–10 relevant chunks
Mid-range	GPT-4.1	1,000,000 tokens [22]	Auto-injected summaries; persistent memory	Web search, code interpreter, API supports custom RAG
Largest advertised (theoretical)	Llama 4 Scout	10,000,000 tokens, theoretical; trained at 256 K [19]	Stateless base; implementation-dependent	No native RAG; developer-implemented via LangChain, LlamaIndex; typical 3–10 chunk retrieval

that addresses use-phase epistemic exclusion as a lifecycle-governable risk with the structural specificity the IIKN framework provides. This review followed no documented systematic review protocol (e.g., PRISMA); it is reported here as a targeted, purposive review, and the novelty claim below should be read accordingly. Existing constructs, as documented in Table 2, address either pre-deployment conditions, such as training bias and model drift, or post hoc evaluation of outputs, such as hallucination detection and audit review. The closest adjacent body of work is the empirical literature on LLM output homogenization, diversity, and mode collapse, as well as on human-AI homogenization of ideas, which demonstrates that AI-assisted writing and ideation measurably reduce content and idea diversity at the population level [23, 1, 9]. That literature establishes the empirical mechanism IIKN presupposes: that ranked, alignment-tuned generation narrows the output distribution. It does not, however, offer a lifecycle-governable construct. Its unit of analysis is aggregate output diversity across many generations, sessions, or users, assessed after the fact, rather than the epistemic scope of a single decision-maker's interaction trajectory, assessed prospectively against a declared reference. IIKN's contribution is to reframe this narrowing dynamic as a use-phase risk object with a declared scope, lifecycle placement, and standards-aligned control set, not to claim the narrowing mechanism itself as new. None of the reviewed governance constructs, including the homogenization literature, renders the interaction-level process by which the epistemic field available for output generation is progressively constrained during live system use auditable against a declared scope. This absence is offered as formative, not exhaustive, evidence for introducing IIKN as a distinct risk category requiring its own governance treatment.

The epistemic effects of these mechanisms are compounded by their interactions. AI-mediated decision support rarely consists of a single-turn exchange. It unfolds through iterative prompt-response sequences in which early outputs condition subsequent questions, refinements, and validations. Over time, this path dependence narrows the trajectory of inquiry, producing outputs that are increasingly coherent and internally consistent while the underlying option space contracts [30, 16].

This paper reframes inference-time interaction as a governance-relevant surface rather than a neutral conduit. The risk is not that systems produce false or misleading information, but that plausible, defensible, and potentially consequential alternatives are never surfaced for consideration. Without explicit epistemic controls, organizations face a

Table 2 Distinguishing IIKN from adjacent epistemic and interaction constructs

Concept	Primary focus	Key distinction from IIKN
Epistemic Closure	Logical insulation of belief systems	Cognitive and philosophical; not lifecycle-operational
Framing Effects	Influence of information presentation	Operates after the informational scope is defined
Agenda-Setting	Prioritization of topics or issues	Assumes visibility of alternatives
Omission Bias	Preference for inaction or missing options	Lacks system-level traceability
Path Dependence	Constraint via early decisions	Human-aware and non-instrumented
IIKN	Interaction-level epistemic exclusion	AI-mediated, opaque, lifecycle-governable
Output Homogenization/Diversity Collapse	Reduction in aggregate content or idea diversity from repeated AI-assisted generation [23, 1, 9]	Empirical and population-level; measures aggregate output similarity across sessions or users rather than auditable exclusion within a single declared-scope interaction

governance condition in which AI-mediated decisions may be defensible and aligned yet systematically incomplete in ways not discoverable through standard assurance reviews.

To name and govern this phenomenon, this paper introduces Interaction-Induced Knowledge Narrowing (IIKN): the systematic narrowing of declared epistemic scope during the use of AI systems, driven by bounded context windows, ranked probabilistic synthesis, and interaction path dependence. The defining characteristic of IIKN is unobserved exclusion rather than error. IIKN is framed as a use-phase risk analytically distinct from training data bias, retrieval quality, model drift, or alignment failure, although it may interact with those risks in practice. Addressing IIKN does not require modifying model weights, expanding datasets, or introducing new performance metrics. It requires governing the design, monitoring, and bounding of interaction and inference-time behaviors within organizational decision contexts.

This paper is conceptual and governance-oriented rather than empirical. Its objective is to articulate a previously underspecified lifecycle risk and establish the definitional and structural foundation required for subsequent empirical research. Empirical validation of prevalence, quantification of impact, and evaluation of control effectiveness are identified as necessary future work, enabled by a clearer definition of risk. The paper addresses both scholars examining the ethical implications of AI-mediated decision-making and practitioners governing AI systems in operational contexts.

2 Formal definition of the risk

For IIKN to be governable, auditable, and operationally actionable, it must be defined with precision and clearly differentiated from adjacent AI risk categories. Without a shared definition, the phenomenon remains anecdotal and resistant to formal control.

2.1 Authoritative definition

IIKN is a lifecycle AI risk in which bounded context windows and ranked probabilistic synthesis progressively constrain the scope of options, perspectives, or considerations surfaced during system use. Through iterative interaction, early truncated outputs anchor subsequent reasoning paths, producing path dependence and systematic exclusion without explicit acknowledgment or intent [31, 26, 30, 16]. IIKN can occur even when outputs are accurate, aligned, and non-deceptive. The risk is realized when no accountable mechanism exists to declare epistemic scope or to validate whether interaction-level exclusion remains within that declared scope. This condition is structural and independent of model quality: it arises from the absence of governance instrumentation at the use-phase level, not from model error, misalignment, or technical failure.

2.2 Why IIKN is not merely incomplete information

All decision processes operate under bounded rationality, and no decision-support system can enumerate all possibilities [16, 18]. IIKN identifies a governance failure mode that bounded rationality alone does not explain. Traditional notions of incompleteness assume that decision-makers are aware of epistemic limits and exercise judgment accordingly. IIKN arises when interaction-level system properties silently pre-structure the decision space, making exclusion neither visible nor contestable.

IIKN differs from satisficing in three decisive respects. First, exclusion is system-induced rather than a deliberate human tradeoff. Second, exclusion is progressively

amplified through interaction path dependence. Third, exclusion becomes durable through the reuse of AI-generated artifacts, transforming situational narrowing into institutional assumption [28, 3].

2.3 Epistemic scope as a governable object

Epistemic scope is treated not as an abstract ideal of completeness, but as a declared governance object against which interaction-level narrowing can be assessed. The use case provides the accountable unit of analysis: individual interaction turns are too granular for governance, while organization-wide scope declarations are too coarse to be meaningful. A minimal epistemic scope declaration includes the following elements:

- Decision type: Operational, tactical, or strategic, including reversibility and expected decision lifespan.
- Stakeholder horizon: Identification of stakeholder groups potentially affected, including downstream or absent stakeholders.
- Coverage dimensions: Categories of considerations expected to be surfaced, such as regulatory, ethical, technical, or organizational.
- Depth posture: Whether breadth-first exploration is required prior to convergence, or whether depth-first narrowing is acceptable by design.
- Reuse horizon: Whether outputs are ephemeral, decision-bound, or intended for institutional reuse.

These elements are intentionally qualitative. Their governance function is to establish an auditable reference against which epistemic narrowing can be acknowledged, accepted, or escalated.

2.4 Positioning IIKN relative to adjacent epistemic constructs

IIKN intersects with, but is not reducible to, several well-established concepts in epistemology, decision science, and human-computer interaction. Table 2 clarifies these relationships.

Classical epistemic closure describes a condition in which a belief system becomes insulated from disconfirming evidence through logical entailment. IIKN differs in that narrowing arises operationally from interaction design, retrieval constraints, and inference-time selection within AI-mediated systems, without presupposing the user's belief adoption. Framing effects and agenda-setting operate at the level of presentation or prioritization; IIKN operates upstream, shaping the epistemic field itself by determining which options are made available for framing in the first place. The output-homogenization and diversity-collapse literature is the closest empirical relative: it documents population-level reduction in content and idea diversity from AI-assisted generation [23, 1, 9], a mechanism consistent with IIKN's account of ranked probabilistic synthesis, but evaluated across aggregated outputs rather than as an auditable property of a single interaction trajectory. Table 2 situates this relationship explicitly.

IIKN is best understood as a synthesis construct that integrates insights from epistemology and HCI(human-computer interaction) while extending them into an auditable, lifecycle-embedded governance risk. Existing frameworks lack mechanisms to render interaction-level epistemic exclusion observable, attributable, and accountable: precisely the gap this paper addresses.

2.5 2.5 Differentiation from adjacent AI risk categories

For governance and audit purposes, IIKN must be distinguishable from adjacent AI risk categories. Table 3 provides functional differentiation to support risk registration, control selection, and standards-aligned classification. IIKN is not a variant of bias, error, or degradation. It is a distinct use-phase epistemic risk concerned with exclusion relative to the declared scope, rather than with correctness or deviation.

2.6 Necessary conditions for IIKN emergence

IIKN emerges reliably when the following structural conditions are jointly present during system use:

- The system operates under finite context constraints that limit simultaneous consideration of alternatives.
- Outputs are generated through ranked probabilistic synthesis without explicit completeness criteria.
- Interaction is iterative, such that early outputs influence subsequent prompts and reasoning paths.
- No explicit mechanism exists to declare, validate, or contest epistemic coverage.

Individually, none of these conditions is sufficient. When they co-occur, epistemic narrowing becomes a structurally induced outcome of interaction design rather than a function of misuse, poor prompting, or model error. Critically, IIKN does not imply malicious intent, deception, or technical malfunction. It is a byproduct of design tradeoffs that optimize fluency, relevance, and efficiency, which explains why it is frequently invisible to conventional assurance processes.

Table 3 Differentiation of IIKN from adjacent AI risk categories

Risk category	Primary concern	Failure signal	How it differs from IIKN
Training Data Bias	Skewed or discriminatory representation learned during training	Disparate outcomes, protected-class impact	Bias concerns who is represented. IIKN concerns which plausible alternatives are excluded during use, regardless of training balance.
Hallucination	Generation of false or unsupported content	Factual errors, unverifiable claims	Hallucination concerns incorrectness. IIKN occurs even when outputs are accurate and well-supported.
Model Drift	Change in model behavior over time	Performance degradation	Drift is temporal. IIKN can manifest immediately in static, well-calibrated models.
Retrieval Failure	Failure to retrieve relevant information	Missing retrieved context	Retrieval failure is a recall defect: relevant information exists and is accessible, but is not returned due to indexing, embedding, or query formulation gaps, and is typically visible as missing or low-relevance results. IIKN begins downstream of successful retrieval, at the ranking and top-k selection step (Sect. 3): even a complete, correctly retrieved candidate pool is still narrowed by ranking and compression before it reaches the model or the user. Accordingly, retrieval-stage ranking bias is a precursor mechanism that feeds into IIKN rather than IIKN itself; the two risks share an upstream boundary and should be assessed jointly in retrieval-augmented systems.

3 IIKN metrics

Section 2.6 establishes that IIKN emerges when four structural conditions co-occur: bounded context windows, ranked probabilistic synthesis, iterative interaction path dependence, and the absence of explicit epistemic scope controls. Formalizing these conditions enables governance systems to render narrowing observable and threshold-bound. The metrics developed here are not diagnostic completeness tests; they are operationalizations of the four-condition emergence model that support trigger-gate calibration, scope-indicator monitoring, and retrieval-diversity assessment, as specified in Sects. 7 and 8. Governance instrumentation requires measurable proxy signals that approximate epistemic narrowing without presupposing ground-truth knowledge of what was excluded. The formalization that follows is offered as an illustrative, not diagnostic, formalization: it renders the emergence conditions in set-theoretic and information-theoretic terms to support governance reasoning and threshold design, not to certify that narrowing has or has not occurred in a given deployment. Two objects central to the formalization, the full hypothesis space and the true theme distribution, are not observable in live deployment; this constraint is foregrounded before the notation is introduced.

In Fig. 1, the user query Q_t defines the initial hypothesis space H_t; ranking and top-k selection produce the evidence set E_t (retrieval narrowing); compression produces the retained context C_t (contextual narrowing); the resulting answer shapes the follow-up query Q_{t+1}, compounding narrowing across steps.

The formalization operates over two measurable narrowing mechanisms that together instantiate three of the four emergence conditions from Sect. 2.6. Retrieval narrowing captures the divergence between the full relevant corpus and the retrieved subset used for generation, operationalizing the bounded-context and ranked-probabilistic-synthesis conditions. Contextual narrowing captures the progressive loss of previously available alternatives through context compression and salience-weighted retention, operationalizing the iterative path-dependence condition. Figure 1 illustrates both mechanisms across interaction steps. Notation is fixed as follows and used consistently throughout this section.

- Let H denote the full hypothesis space: the complete set of themes, perspectives, and explanations relevant to the use case. H is a theoretical reference object; it is not observable or enumerable in live deployment, and no estimation procedure in this paper claims to recover it exactly.
- Let $H_t \subseteq$ H denote the subset of themes actively represented in system state at step t, and $|H_t|$ its cardinality, the count of distinct themes represented at step t;

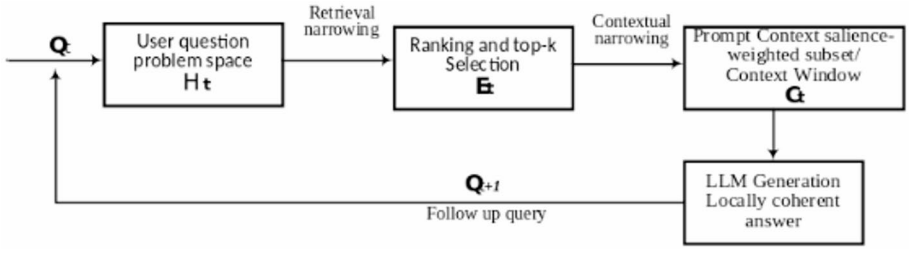


Fig. 1 Retrieval narrowing and contextual narrowing across iterative interaction steps

the cardinality operator $|\cdot|$ is used whenever a count, rather than the set itself, is intended.

- Let $E_t \subseteq \mathrm{H}$ denote the evidence set surfaced by retrieval at step t, and C_t the content retained in the context window at step t after compression.
- Let S_t denote the empirical theme distribution induced by E_t: the proportion of retrieved or represented items assigned to each theme in Themes (E_t), normalized to sum to one. S_t is therefore always a distribution derived from a set, never the set itself, resolving the set/distribution ambiguity in the metrics below.
- Let P denote the true theme distribution over Themes (H): the reference distribution against which narrowing is assessed. Like H, P is not observable in live deployment; Sect. 3's final paragraph addresses how it is approximated.

Given a user query Q_t, the system retrieves a candidate pool from a corpus using vector similarity or keyword search, then narrows this pool to the k most relevant documents; this is retrieval narrowing. Once the top-k documents are combined into a prompt, a compression step retains salient tokens or sentences and discards the remainder; this is contextual narrowing. The locally coherent answer at step t influences the follow-up query Q_{t+1}.

In retrieval-augmented systems, IIKN is therefore not solely a generation artifact; it is introduced upstream through the evidentiary field that ranked the retrieval constructs. In agentic architectures, the same dynamic operates across the delegation and coordination steps, such that alternatives not explicitly carried forward through intermediate task states are effectively excluded from the epistemic space available to downstream agents. Section 4 develops these boundary conditions in detail.

IIKN occurs when $|H_{t+1}| < |H_t|$ and $H_{t+1} \not\supseteq H_t$: the active hypothesis set both shrink in count and lose members it previously contained, rather than merely being relabeled or reordered. Thount comparison alone ($|H_{t+1}| < |H_t|$) is necessary but not sufficient to establish narrowing in the sense this paper cares about, since a smaller set that happens to contain entirely different, equally broad themes would satisfy the inequality without representing exclusion; the non-superset condition ($H_{t+1} \not\supseteq H_t$) is what establishes that specific, previously available themes were dropped.

Let $E_t = \mathrm{TopK}(R(Q_t))$, where R is a retriever, and Q_t is the query. Retrieval narrowing arises because $E_t \subset H$, and selection is biased by ranking and corpus structure. This boundary is definitional, not incidental: retrieval failure is assessed against what a complete and correctly functioning retriever would have returned, while IIKN is assessed against what a successful retrieval made available, but ranking, top-k truncation, or compression subsequently excluded (see Table 3). Systems that combine both risks, for example, a retriever with limited recall feeding a ranker that further narrows an already-incomplete pool, require controls for each addressed separately, since fixing one does not resolve the other.

Due to bounded capacity and salience weighting, the context window at step $t+1$ can be defined as:

$$C_{t+1} = f\left(C_t, \text{ new tokens}\right) \tag{3.1}$$

where f is a lossy compression function (attention + truncation). Thus, contextual narrowing occurs as:

$$\mathcal{H}_{t+1} \subseteq \mathcal{H}_t \cup ?_t \tag{3.2}$$

with salience-weighted retention. Equation 3.2 is a boundedness property, not itself an operational test for narrowing: it states that the next step's represented themes cannot exceed what was already represented plus what retrieval newly surfaced, which holds by construction and does not by itself demonstrate that exclusion occurred. The operational narrowing criteria are the cardinality inequality above and the coverage ratio introduced next; Eq. 3.2 should be read as a scope constraint on H_{t+1} rather than as evidence of narrowing.

Three metrics are proposed as governance-instrumentation primitives. Each operationalizes a distinct dimension of the narrowing condition and maps to a specific governance function: threshold calibration for trigger gates, scope-drift detection, and retrieval-diversity monitoring. None asserts ground-truth completeness; each provides an auditable proxy signal against which declared epistemic scope can be assessed.

(a) Coverage ratio r_t.

Coverage ratio r_t measures the degree of theme shrinkage in step *t*, and is defined as:

$$r_t = |\text{Themes}(\mathcal{H}_t)| \,/\, |\text{Themes}(\mathcal{H})| \tag{3.3}$$

(b) Distribution divergence v_t

Distribution divergence v_t measures how the empirical distribution S_t (defined above as the theme distribution induced by the evidence set E_t) diverges from the reference distribution *P*, using Kullback-Leibler divergence (relative entropy). Because S_t is defined as a function of the observable set E_t, it can be computed directly from logged retrieval output; *P* cannot be observed directly and must be approximated, as addressed below.

$$v_t = D_{KL}(S_t \parallel P) = \sum {}_i S_t(i) \log\left(\frac{S_t(i)}{P(i)}\right) \tag{3.4}$$

(c) Entropy reduction $E(S_t)$ measures the decrease of randomness in hypothesis space at step *t*, and is defined as:

$$E(S_t) = -\sum {}_i S_t(i) \log S_t(i) \tag{3.5}$$

Narrowing implies $E(S_{t+1}) < E(S_t)$

These three metrics operate as complementary governance-instrumentation primitives. Coverage ratio and entropy reduction support early detection of retrieval diversity collapse, providing the proxy signals that trigger the gate thresholds required in Sect. 7.3. Distribution divergence serves as the principal scope-drift indicator: a rising v_t signals that the active retrieval distribution has materially departed from the declared epistemic scope, warranting activation of the coverage confirmation checkpoint as specified in Sect. 8.3. Together, they provide the observability layer that renders IIKN governable at the interaction level rather than recoverable only in retrospect. The fourth emergence

condition from Sect. 2.6, absence of explicit epistemic scope controls, is addressed not through measurement but through the organizational controls in Sect. 8.

A practical operationalization challenge for distribution divergence v_t is estimating the reference distribution P in live-deployment contexts. P is not directly observable, and this limitation should be treated as a standing constraint on the metric rather than a solvable preprocessing step. Two estimation strategies are recommended as primary: (1) domain corpus sampling, in which P is estimated from a stratified sample of the relevant subject-matter literature prior to deployment and used as a stable reference distribution, and (2) expert elicitation, in which domain specialists enumerate expected theme categories and assign prior probability weights as part of the epistemic scope declaration process. A third strategy, multi-session baseline comparison, approximates P from the theme distribution observed across a large sample of initial-turn responses to structurally similar queries. This strategy is flagged explicitly as weaker than the first two: it treats first-turn output as a distributional baseline, but Sect. 3's own account of path dependence holds that first-turn output is already shaped by ranked probabilistic synthesis and may itself be narrowed relative to H. Using it as a ground-truth proxy is therefore partially circular. Where no corpus sample or expert elicitation is feasible, multi-session comparison may still be used as a conservative, lower-bound estimate. Still, governance systems should document this choice explicitly and treat resulting v_t values as indicating a floor on divergence, not its true magnitude. Every estimation strategy introduces error; governance systems should document the chosen method as part of the epistemic scope declaration and treat v_t thresholds as calibration targets subject to periodic review rather than fixed operational constants.

Section 3, as a whole, should be read as a formalization proposal rather than as a validated measurement instrument. None of the three metrics, nor the estimation strategies above, has been operationalized against a deployed system, a reconstructed interaction log, or even a synthetic benchmark in this paper; their behavior under realistic noise, adversarial gaming, or estimation error is untested. Empirical validation of the metrics themselves, not only of IIKN's prevalence, is accordingly a precondition for any organization adopting them as operational thresholds rather than illustrative constructs and is included among the empirical research priorities identified in the Limitations section.

4 Mechanisms of emergence and boundary conditions

IIKN does not arise from a single defect, misuse pattern, or design error. It emerges predictably from the interaction of multiple system properties that govern how information is selected, retained, and synthesized during inference. When these properties operate together, epistemic narrowing is structurally induced, particularly in the absence of explicit epistemic scope controls. The objective here is governance clarity: to demonstrate why IIKN cannot be adequately addressed through accuracy testing, bias mitigation, or post hoc audit alone.

IIKN can be introduced through various ways:

- *Contextual selection layers*: Contemporary LLM-based systems apply selection pressure at multiple contextual layers prior to output generation: prompt interpretation, retrieval filtering, relevance ranking, summarization, alignment tuning, and response compression. Each layer independently reduces epistemic scope; collectively, they determine which informational alternatives are rendered

visible to the user. These selection processes are probabilistic and opaque. Users encounter outputs that appear comprehensive while silently omitting viable perspectives, interpretations, or solution pathways.

- *Inference-time dynamics and epistemic exclusion*: IIKN is fundamentally an inference-time phenomenon. Prompt interpretation operates as an initial scope filter, translating underspecified queries into a constrained semantic frame. Retrieval and ranking mechanisms select passages based on similarity and relevance scores rather than representativeness [17, 14], which can privilege high-salience or majority-aligned sources at the expense of minority, dissenting, or edge-case perspectives. This selection dynamic is structurally consistent with algorithmic curation effects documented in the echo-chamber literature [8]. Summarization and compression further collapse epistemic diversity by privileging coherence and fluency over pluralism. These processes are individually defensible from an engineering perspective, but their interaction produces a cumulative narrowing effect that is rarely instrumented or governed.
- *Human-in-the-loop interaction dynamics*: In human-mediated interactions, IIKN often emerges through iterative confirmation dynamics. Users tend to accept fluent outputs as exhaustive, reinforcing the system's inferred intent through follow-up prompts that further constrain scope [24, 10]. Alignment tuning may amplify this effect by optimizing for user satisfaction rather than epistemic breadth. IIKN manifests as a gradual contraction of the informational field, often unnoticed by users who lack visibility into excluded alternatives. Human agency remains present but is exercised within an already-narrowed epistemic space.
- *Boundary conditions in agentic and autonomous systems*: IIKN risk intensifies markedly in agentic and semi-autonomous systems. In these configurations, epistemic narrowing may occur without direct user confirmation, driven by architectural features such as memory compression, tool-routing heuristics, intermediate summarization, and objective-driven pruning. Unlike conversational systems, agentic architectures often abstract or discard interaction histories, rendering epistemic exclusions invisible to both users and auditors. In multi-agent systems, task decomposition and delegation further fragment epistemic visibility. Whereas human-in-the-loop IIKN may be mitigated through interface design, agentic IIKN requires explicit lifecycle controls, including instrumentation of epistemic scope, interaction logging, and trigger-based escalation.

Because IIKN operates through omission rather than commission, downstream audit is structurally insufficient. Auditing outputs for accuracy, bias, or compliance cannot reveal which alternatives were never surfaced. Post hoc review faces three insurmountable limitations: absence leaves no evidentiary trace; once interaction histories are compressed or discarded, the epistemic trajectory cannot be recovered; and counterfactual analysis requires knowing what could have been surfaced, which is categorically different from evaluating what was produced. This structural limitation necessitates a shift toward interaction-aware lifecycle controls. IIKN must be rendered observable at the point of emergence, not inferred retrospectively.

5 Illustrative interaction dynamics

The following scenarios are hypothetical illustrative constructions, not accounts of documented, specific cases. They are built to be consistent with patterns reported in the general governance, human-computer interaction, and automation-reliance literature cited throughout this paper. Still, no single published case is claimed to correspond to either scenario, and they do not assert statistical prevalence or empirical generalization. Their purpose is to demonstrate why IIKN is difficult to detect through conventional assurance mechanisms and why governance intervention must occur upstream rather than through artifact review alone.

Scenario 1: Policy design support

An organization uses an AI system to support the development of an internal AI governance policy, a use pattern consistent with patterns described in the governance and information systems literature [12, 32]. However, the scenario itself is a hypothetical construction rather than a specific reported case. A decision-maker prompts the system to identify key considerations for deploying a generative AI capability in a regulated environment. The system surfaces commonly cited themes: data privacy, transparency, human oversight, and regulatory compliance. These are accurate and widely recognized, reflecting the dominant risk categories codified in ISO/IEC 42,001 and the NIST AI RMF. Less frequently discussed considerations, such as long-term epistemic degradation, interaction-level framing effects, or institutional dependence on AI-generated artifacts, are not brought to light.

The user proceeds with follow-up questions that refine and operationalize the initial themes. Subsequent prompts focus on implementation guidance and audit readiness. The system introduces no additional governance dimensions. The resulting policy draft appears comprehensive and defensible. Yet, entire classes of governance risk were never considered because they were excluded from the epistemic scope established in the first interaction turn.

Scenario 2: Architectural design review

An engineering team prompts an AI system to recommend architectural patterns for integrating a large language model into an existing platform, a task consistent with patterns of AI-assisted technical review described in the human-automation interaction and trust literature [24, 10]. However, the scenario itself is a hypothetical construction rather than a specific reported case. The system surfaces widely adopted patterns involving API gateways, retrieval-augmented generation, logging, and monitoring. Follow-up questions focus on performance optimization, scalability, and cost management. Alternative approaches, including deliberately constrained deployments, non-generative decision-support hybrids, or architectures prioritizing epistemic auditability over throughput, are never surfaced. These alternatives are not invalid, but they are less canonical in the training distribution and therefore less likely to emerge through ranked probabilistic synthesis in early interaction turns, consistent with the retrieval bias mechanism formalized in Sect. 3.

In both scenarios, narrowing arises through the interaction of mechanisms described in Sect. 3: ranked probabilistic synthesis privileges widely documented options in initial responses, truncated enumeration presents a limited set of considerations without signaling incompleteness, and anchoring effects from initial outputs shape subsequent prompts. No individual step appears deficient. Each response is accurate, relevant, and

aligned with user intent. The narrowing emerges cumulatively across interaction rather than as a discrete failure point. At no point is the user informed of what was not surfaced, and the absence of additional options is structurally positioned to be interpreted as nonexistence rather than as a consequence of inference-time selection, a dynamic consistent with documented automation bias and over-reliance effects in human-AI collaborative decision support [24, 13].

6 Lifecycle risk mapping

To be governable, IIKN must be treated as a lifecycle-native risk rather than a situational artifact of individual interactions. Framing IIKN as a conversational or prompt-level issue obscures where epistemic exclusion is introduced, how it is amplified, and why late-stage controls routinely fail to detect it. The analysis here emphasizes a central governance insight: once introduced, epistemic exclusion is only partially reversible. At each lifecycle phase, decisions about scope, interaction structure, and reuse progressively eliminate visibility into alternatives that were never surfaced.

6.1 Use case definition

IIKN is often introduced before any system is deployed. During use-case definition, organizations implicitly define what constitutes an acceptable AI-assisted output without explicitly declaring the epistemic scope. Use cases frequently emphasize efficiency and decision velocity while leaving assumptions about representativeness and breadth unstated. When scope expectations are implicit, AI systems are authorized by default to optimize for relevance and fluency over epistemic coverage. Without an explicit reference point, downstream controls lack a basis for evaluating whether narrowing has occurred or remains acceptable.

6.2 Architecture and interaction design

IIKN becomes structurally embedded during architecture and interaction design. Decisions regarding context window allocation, default response length, summarization behavior, retrieval configuration, and interaction flow determine how epistemic scope is constrained during use. Context windows function as epistemic selection surfaces rather than neutral performance constraints. Interaction designs that emphasize concise outputs, ranked recommendations, or single-path reasoning normalize early convergence and increase the likelihood of unacknowledged exclusion. Once embedded at this stage, epistemic exclusion cannot be reliably addressed through review or monitoring alone.

6.3 Deployment and operational use

During live deployment, IIKN is actively amplified through iterative interaction. Early outputs anchor subsequent prompts, refinements, and validations. Operational conditions intensify this effect: time pressure, cognitive load, and organizational incentives favor convergence over reconsideration. The system's fluency reinforces user confidence, reducing the likelihood that unpresented alternatives are actively sought. At this phase, IIKN becomes behaviorally reinforced through human-AI co-adaptation.

6.4 Monitoring and audit

Most existing monitoring and audit mechanisms are structurally ill-suited to detect IIKN. Output accuracy checks, bias metrics, and performance indicators evaluate produced content rather than excluded alternatives. Because IIKN manifests as an absence rather than a deviation, it frequently goes undetected. Once interaction histories are compressed, summarized, or discarded, the epistemic trajectory that produced a given output cannot be reconstructed without specialized logging.

6.5 Feedback and reuse

The most consequential amplification of IIKN occurs when AI-generated outputs are reused beyond their original interaction context. Recommendations are incorporated into policies, architectures, standards, or institutional knowledge repositories. Reuse strips outputs of their interaction context. Downstream consumers encounter conclusions without visibility into what alternatives were never surfaced, creating epistemic echo effects that reinforce dominant frames across decisions, teams, and organizational cycles. At this point, IIKN transitions from a situational interaction effect to a systemic organizational risk.

7 Real-time observability and lifecycle governance

IIKN cannot be governed solely through outcome-based review or periodic audit. Because IIKN manifests through omission at inference time, governance controls must operate continuously, at the interaction level, and in alignment with system lifecycle stages. This section specifies how IIKN can be rendered observable and actionable during system use.

7.1 Requirements for IIKN observability

Effective governance of IIKN requires epistemic scope to be made partially observable during interaction without exposing proprietary model internals. Three requirements are foundational. First, interaction events must be logged with sufficient granularity to support scope analysis, including prompt intent interpretation, retrieval filters applied, and summarization or compression steps invoked. Second, epistemic scope indicators must be defined relative to the use context: what constitutes unacceptable narrowing in exploratory research differs from what constitutes unacceptable narrowing in high-stakes decision support. Third, observability must be continuous rather than episodic. IIKN can emerge within a single interaction or cascade across turns. This instrumentation does not contradict the earlier claim that IIKN is undetectable through standard output-focused assurance. That claim concerns retrospective detection: auditing a finished output for accuracy, bias, or compliance cannot reveal what was never surfaced. The proxy signals developed in Sect. 3 and operationalized here instead support prospective observability: they flag conditions correlated with narrowing as the interaction unfolds, at the point of emergence, rather than reconstructing, after the fact, what was excluded. Under this framework, detection means observing elevated narrowing risk in real time, not verifying completeness after output is produced.

7.2 Epistemic scope indicators and proxy signals

Because epistemic scope cannot be measured directly, governance systems must rely on proxy indicators. These do not assert the completeness of ground truth; they signal when interaction dynamics are likely to exclude viable alternatives. Examples include sudden reductions in semantic variance across turns, repeated reinforcement of a single interpretive frame, aggressive summarization under uncertainty, or consistent suppression of low-salience but relevant sources. Such signals are particularly important in retrieval-augmented and agentic systems, where routing and pruning decisions may occur without user awareness.

Figure 2 shows epistemic scope proxy indicators across interaction stages. Measurable signals, including reduced semantic variance, repeated framing, source suppression, and aggressive summarization, emerge during retrieval, generation, and multi-turn interaction. Their convergence signals elevated IIKN risk and justified governance intervention. Percentages shown are illustrative values used to depict convergence across stages, not measured empirical frequencies.

7.3 Trigger gates and proportionate intervention

Trigger gates are a specialization of established stage-gate practices widely used in enterprise governance. What distinguishes IIKN trigger gates is their focus: traditional lifecycle gates evaluate readiness, compliance, or performance; IIKN trigger gates evaluate epistemic accountability, whether the narrowing introduced during AI-mediated interaction aligns with declared scope and organizational intent. Trigger gates are particularly well-suited to IIKN governance because epistemic exclusion is often irreversible once interaction trajectories have converged or outputs have been reused. Trigger gates intervene at moments when epistemic scope remains contestable.

To avoid overcorrection or unnecessary friction, IIKN governance must be threshold-based. Trigger gates provide a mechanism for escalating oversight only when epistemic narrowing exceeds predefined tolerances. Trigger activation does not imply system

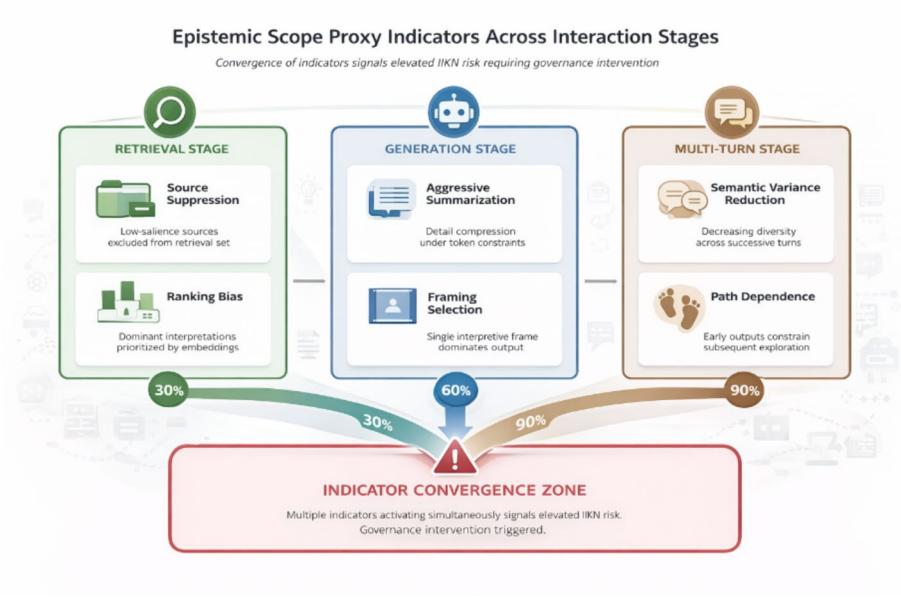


Fig. 2 Representative epistemic scope proxy indicators across interaction stages

Table 4 Mapping IIKN Observability Mechanisms to AI Lifecycle Stages and Governance Responsibilities

Lifecycle stage	IIKN observability mechanisms	Primary governance purpose	Accountable roles
Use Case Design & Authorization	Epistemic scope assumptions; initial framing constraints; expected alternative space	Prevent structurally narrow problem definitions; document intended epistemic breadth.	Product owner; domain lead; AI governance committee
System Design & Architecture	Interaction logging schema; retrieval and routing visibility; summarization thresholds	Ensure architectural choices do not preclude epistemic diversity by design.	AI architects, governance, and risk leads
Model Integration & Configuration	Proxy indicators (semantic variance, framing reinforcement); trigger threshold calibration	Detect early epistemic contraction during inference-time operation	ML engineers; model risk management
Deployment & Live Operation	Real-time interaction monitoring; trigger gate activation logs; scope alerts	Enable timely intervention before epistemic narrowing propagates downstream.	Operations teams; human-in-the-loop reviewers
Monitoring & Audit	Reconstructable interaction traces; scope signal histories; escalation records.	Support internal accountability, external audit, and regulatory inquiry	Internal audit; compliance; regulators
Model Update & Retirement	Longitudinal IIKN trend analysis; cumulative narrowing assessments	Inform retraining decisions, decommissioning, or scope redesign	AI governance board; executive sponsors

failure; it signals a heightened epistemic risk that requires attention. This approach aligns with specific requirements in both governance standards. ISO/IEC 42001:2023 Clause 6.1 requires organizations to determine actions to address AI risks proportionate to their potential impact; Clause 9.1 mandates monitoring, measurement, and evaluation of the AI management system at planned intervals. IIKN trigger gates can be used to address both requirements at the interaction level. Within the NIST AI RMF, the GOVERN function establishes accountability for AI risk; the MEASURE function calls for ongoing monitoring of AI system behavior in deployment; and the MANAGE function specifies that identified risks be treated, tracked, and reported. IIKN interaction logging and proxy signal monitoring can facilitate to fulfill the MEASURE function requirements; trigger gate activation and escalation records can facilitate to fulfill the MANAGE function reporting and accountability requirements. ISO/IEC 42,001 Clauses 6.1 and 9.1 and the NIST GOVERN, MEASURE, and MANAGE functions require risk treatment and monitoring in general terms; they do not name epistemic exclusion as a monitored condition. IIKN governance is accordingly best framed as consistent with and operationalizable within those existing requirements, rather than as a use-phase instantiation of controls that the frameworks already require in specific form [15, 21]. This is a narrower and more defensible claim: it positions IIKN controls as one reasonable way to satisfy general monitoring and risk-treatment obligations for a risk category that those frameworks do not yet name explicitly, rather than asserting that the frameworks already mandate this specific instrumentation.

7.4 Lifecycle integration and traceability

IIKN governance must be embedded across the AI system lifecycle rather than confined to deployment. Interaction-level observability feeds into design reviews, use-case authorization, model updates, and retirement decisions. When IIKN-related triggers are activated, organizations must be able to reconstruct how epistemic scope was constrained during the interaction, thereby supporting internal accountability, external audits, and regulatory inquiries without requiring disclosure of proprietary model internals. Table 4

maps IIKN observability mechanisms to lifecycle stages and corresponding governance responsibilities.

7.5 Applicability to agentic and multi-agent systems

In agentic and multi-agent architectures, real-time observability becomes critical. Epistemic narrowing may occur entirely within internal planning, delegation, or memory abstraction processes. Without explicit instrumentation, such narrowing remains invisible until it manifests as a downstream error. Agentic systems require additional safeguards, including state-transition logging, bounded autonomy constraints, and explicit scope checkpoints before action execution.

8 Governance controls and design patterns

Mitigating IIKN does not require exhaustive enumeration, encyclopedic outputs, or intrusive technical instrumentation. The governance objective is accountability for epistemic exclusion: ensuring that when narrowing occurs, it is intentional, proportionate, and visible to accountable decision-makers. The controls defined here are designed to integrate into existing governance, risk, and lifecycle review processes rather than to create parallel oversight structures.

8.1 Epistemic scope declaration as a governable object

The foundational control for IIKN governance is the explicit declaration of epistemic scope at the use-case level. A minimal Epistemic Scope Object consists of five fields: Decision Type (operational, tactical, or strategic, including reversibility and expected lifespan); Stakeholder Exposure (internal and external stakeholders potentially affected); Coverage Dimensions (categories of alternatives, risks, or perspectives expected to be surfaced); Depth Expectation (whether breadth-first exploration is required prior to convergence); and Reuse Horizon (whether outputs are ephemeral, decision-bound, or intended for institutional reuse). This object does not prescribe which alternatives must be included. It establishes a reference against which interaction-level narrowing can be assessed. In practice, scope declarations can be embedded in existing use-case approval documents, risk assessments, or design briefs. The critical IIKN-specific distinction from standard requirements documentation is the declaration of epistemic expectations prior to interaction, rather than as a post hoc evaluation criterion. Standard use-case documentation specifies what the system should produce; an epistemic scope declaration specifies the boundaries within which epistemic exclusion is acceptable. This distinction is what makes the declaration auditable against interaction-level behavior rather than only against output correctness.

8.2 Breadth-before-depth interaction patterns

Breadth-before-depth patterns require AI systems to surface categories, dimensions, or classes of alternatives before converging on a detailed analysis of any single option. This pattern directly counteracts anchoring effects [30, 18] by delaying early convergence and making the structure of the decision space visible before refinement occurs. Breadth does not imply exhaustiveness; it requires signaling diversity and uncertainty within the declared epistemic scope. These patterns are most appropriate when decisions are strategic, irreversible, or reused; when declared scope spans multiple stakeholder

perspectives; or when early framing effects would materially constrain downstream options. They are unnecessary for tightly bound operational tasks.

8.3 Coverage confirmation checkpoints

Coverage confirmation checkpoints introduce deliberate pauses where epistemic scope alignment is explicitly assessed. For governance purposes, coverage confirmation is satisfied when documented evidence demonstrates that an epistemic scope declaration exists, that surfaced outputs plausibly align with that declared scope, and that any narrowing beyond the declared scope has been acknowledged and either accepted or escalated by an accountable role. Checkpoints do not require enumerating excluded alternatives or proving completeness. Their governance value lies in creating auditable moments where narrowing becomes a conscious decision rather than an implicit default. This distinguishes IIKN coverage confirmation from standard design reviews, which evaluate outputs against functional requirements, and from ethical impact assessments, which evaluate outputs against normative criteria. Coverage confirmation checkpoints evaluate whether the interaction process by which outputs were generated was epistemically bounded in a declared and proportionate way, a procedural accountability standard rather than a substantive one.

8.4 Human-in-the-loop epistemic validation

In high-impact or irreversible decision contexts, human oversight remains essential. Unlike traditional human-in-the-loop controls focused on correctness or safety [24, 13], epistemic validation assesses whether relevant alternatives may have been excluded relative to the declared scope. This role should be performed by individuals with domain and contextual expertise, not by system developers or prompt engineers. The objective is not to override AI outputs, but to interrogate the boundaries of what was considered. Epistemic validation is most appropriate when decisions affect strategy, policy, or long-term commitments; when outputs are intended for institutional reuse; or when stakeholder exposure extends beyond the original interaction context. The effectiveness of this role depends on what is exposed to the reviewer, not merely on a reviewer's presence: research on AI explanations and complementary human-AI team performance shows that oversight quality is contingent on what the interface surfaces for scrutiny, not on human involvement alone [4].

8.5 Provenance and interaction context tagging

When AI-generated outputs are reused beyond their original interaction context, provenance and context tagging preserve visibility into epistemic conditions that would otherwise be lost. This practice extends transparency from model-level to interaction-level artifacts, paralleling established documentation and accountability frameworks such as model cards [20] and internal audit frameworks for algorithmic accountability [27]. At a minimum, reused artifacts should indicate that the content represents a synthesized subset rather than an exhaustive analysis, the epistemic scope under which it was produced, and any assumptions or usage limitations associated with the original context.

8.6 Proportionality and control calibration

IIKN controls must be calibrated to decision impact and reuse potential rather than applied uniformly. Controls are generally unnecessary when the task is low-impact and reversible, outputs are ephemeral and not reused, or external rules tightly constrain the decision space. Explicitly documenting these exclusions strengthens audit defensibility by demonstrating intentional governance rather than omission. The objective of IIKN governance is not to eliminate narrowing, but to govern it. By embedding epistemic scope declaration, coverage confirmation, and proportional validation into existing governance workflows, organizations can materially reduce IIKN risk while preserving the efficiency and usability that make AI systems valuable.

9 Conclusion

Epistemic narrowing during AI system use is not a peripheral concern, a user-experience artifact, or a prompt-engineering nuance. It is a governance-relevant lifecycle risk with material implications for decision quality, accountability, and institutional resilience. IIKN names a failure mode that existing AI governance frameworks implicitly assume away: that fluent, accurate, and aligned outputs may nonetheless be systematically incomplete in consequential ways.

The strategic significance of IIKN lies in its invisibility. Unlike bias, error, or drift, IIKN manifests as absence rather than deviation: it produces decisions that are defensible, coherent, and professionally structured while quietly excluding alternatives that were never surfaced, evaluated, or challenged. For boards and senior leaders, the implication is direct. Unmanaged IIKN typically produces no failure signal; it produces confident, plausible, and incomplete decisions that are difficult to contest once institutionalized.

The governance controls proposed here, including epistemic scope declaration, breadth-before-depth interaction patterns, coverage confirmation checkpoints, and lifecycle-aligned trigger gates, are proportional rather than exhaustive. Their relevance will intensify as organizations adopt agentic and multi-agent systems: as interaction chains lengthen, delegation increases, and intermediate decisions are made autonomously, epistemic narrowing can compound faster and propagate further than traditional oversight mechanisms are designed to detect. Treating interaction design as governance infrastructure is not optional in these environments; it is a prerequisite for responsible scale.

Ultimately, governing AI systems requires attention not only to what they produce but also to what they structurally exclude in use. The absence of error should not be mistaken for the presence of completeness. Recognizing and managing IIKN is therefore not an enhancement to AI governance, but a necessary condition for AI systems that support well-considered, accountable, and resilient decision-making.

It is worth noting that epistemic narrowing is not inherently undesirable. In many contexts, focused, convergent, and efficient interaction is precisely what users need. Operational tasks, tightly bound technical queries, and exploratory ideation often benefit from systems optimized for relevance and coherence over exhaustive coverage. The objective of IIKN governance is not to eliminate narrowing but to make it explicit, proportionate, and accountable where exclusion carries material consequences. Governance maturity is demonstrated as much by the disciplined choice not to apply IIKN controls in low-risk contexts as by their rigorous application in high-stakes decisions.

This paper introduces IIKN as a governance-relevant lifecycle risk and proposes controls to render epistemic exclusion accountable. The scope and limitations of this work must be clearly acknowledged.

- *Absence of empirical validation*: This paper does not provide empirical validation of the prevalence, magnitude, or frequency of IIKN across organizational contexts. The mechanisms described in Sect. 4 reflect documented structural properties of contemporary LLM-based systems established in the transformer architecture and RAG literature [17, 14, 31, 26]; the governance failure modes identified represent risk projections grounded in that structural evidence rather than organizational observational data. The literature gap analysis conducted for this paper found no existing governance construct that addresses use-phase epistemic exclusion with the specificity provided by the IIKN framework, constituting formative evidence of the gap the paper addresses. The lack of organizational prevalence data and controlled intervention studies is a deliberate constraint of scope. The paper's contribution is conceptual and governance-oriented: it names a previously uncharacterized risk, differentiates it from adjacent categories, explains its structural mechanism, and proposes proportionate operational controls grounded in existing standards requirements. Three empirical research programs are identified as the highest priority for follow-on work: (1) a controlled interaction study measuring semantic diversity contraction across breadth-first versus unstructured interaction conditions; (2) a corpus analysis of AI-assisted organizational documents measuring coverage alignment against declared epistemic scope fields; and (3) a structured practitioner survey assessing recognition of and current organizational response to IIKN in operational AI governance programs. Empirical research to validate prevalence, measure organizational impact, and test control effectiveness represents essential future work. Organizations implementing IIKN governance should treat proposed controls as hypotheses to be validated through piloting, monitoring, and iterative refinement rather than as empirically proven interventions.
- *Controls as conceptual frameworks requiring adaptation*: The governance controls proposed in Sect. 8 are intentionally high-level to accommodate organizational diversity in governance maturity, risk appetite, and operational context. They function as conceptual frameworks rather than prescriptive procedures. Epistemic scope declarations, breadth-before-depth patterns, coverage checkpoints, and trigger gates will manifest differently across industries, use cases, and organizational cultures. Implementation will require piloting, stakeholder engagement, and iterative refinement. This gap between designed frameworks and practiced implementation is not unique to IIKN governance; comparable divergence between imagined and practiced AI development processes has been documented in specialized domains such as healthcare [6].
- *Focus on LLM-based interactive systems*: This paper's analysis focuses on LLM-based systems that mediate decision-making through iterative human-AI interaction. IIKN governance may not generalize directly to fully automated AI systems, non-generative AI systems such as classification or recommender models, emerging architectures that fundamentally depart from transformer-based attention, or hybrid workflows in which AI serves as a supplementary tool rather than a primary information source. Organizations deploying non-LLM AI systems may find

elements of IIKN governance applicable, particularly epistemic scope declaration and reuse controls. Still, they should not assume all mechanisms and failure modes generalize without adaptation.

- *Governance framework, not technical solution*: IIKN governance is a risk management framework, not a technical intervention. It does not propose modifications to model architectures, training procedures, or inference algorithms. Technical modifications, such as expanding context windows or altering ranking algorithms, may shift where epistemic selection occurs, but do not eliminate the fundamental governance question: how should organizations account for what their AI systems exclude in use? The controls operate at the level of interaction design, organizational process, and accountability assignment.

Appendix A: IIKN risk register entry template

Risk name: Interaction-Induced Knowledge Narrowing (IIKN).

Risk category: Use-Phase Epistemic Risk.

Lifecycle phase(s): Use Case Definition; Interaction and Architecture Design; Deployment and Operational Use; Monitoring and Audit; Feedback and Reuse.

Risk description: The risk is that bounded context windows, ranked probabilistic synthesis, and iterative interaction dynamics systematically constrain the set of options, sources, or perspectives surfaced during AI system use, resulting in unobserved exclusion even when outputs are accurate and aligned.

Potential impacts: Incomplete strategic, architectural, or policy decisions; overconfidence in AI-assisted recommendations; institutionalization of narrowed assumptions through reuse; reduced adaptive capacity and resilience.

Inherent risk rating: Medium to High (context-dependent).

Key Controls: Epistemic scope declaration; breadth-before-depth interaction patterns; coverage confirmation checkpoints; human epistemic validation; trigger gate enforcement.

Residual risk considerations: Residual risk remains when narrowing is accepted by design or is proportional to the decision's impact.

Appendix B: Epistemic coverage validation checklist

This checklist supports coverage confirmation checkpoints and IIKN trigger gate evaluations. It is designed to validate alignment between interaction-level outputs and a declared epistemic scope without requiring exhaustive enumeration or reconstruction of excluded alternatives.

- Has an epistemic scope been explicitly declared for this use case?
- Is the decision type (operational, tactical, strategic) and reversibility documented?
- Are the relevant stakeholder groups identified?
- Are the expected coverage dimensions specified?
- Did the interaction follow the declared depth posture (breadth-first or depth-first)?
- Were any assumptions about completeness made explicit during interaction?
- Was epistemic narrowing beyond the declared scope acknowledged?
- If narrowing exceeded the scope, was it accepted or escalated by an accountable role?
- Is the output intended for reuse beyond the immediate interaction?
- If reused, has epistemic scope or provenance context been preserved?

Appendix C: Sample interaction audit questions

- What alternatives were explicitly surfaced during interaction?
- What alternatives were implicitly excluded, and why?
- How did early outputs influence subsequent prompts or decisions?
- Was epistemic narrowing aligned with the declared use case scope?
- Could reasonable stakeholders challenge the completeness of the output based on available artifacts?

The audit focus should remain on process evidence rather than on reconstructing exhaustive option sets.

Appendix D: Glossary entry

Interaction-Induced Knowledge Narrowing (IIKN): A lifecycle AI risk in which bounded context windows and ranked probabilistic synthesis mechanisms progressively constrain the set of options, sources, or perspectives surfaced during system use. IIKN arises through iterative interaction path dependence and can occur even when outputs are accurate, aligned, and non-deceptive. The defining characteristic of IIKN is unobserved exclusion rather than error.

Acknowledgements
Not applicable.

Author contributions
Dale Rutherford and Ningning Wu wrote the main manuscript text. All authors reviewed the manuscript.

Funding
The authors declare that no funds, grants, or other support were received for this study.

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
The authors give the consent for publication of this research.

Competing interests
The authors declare no competing interests.

Received: 6 May 2026 / Accepted: 21 August 2026



References

1. Anderson BR, Shah JH, Kreminski M. Homogenization effects of large language models on human creative ideation. In Proceedings of the 16th Conference on Creativity and Cognition (C&C '24). ACM; 2024. pp. 1–13. https://doi.org/10.1145/3635636.3656204.
2. Anthropic. Claude Opus 4.5 model documentation. 2025. https://docs.claude.com/en/docs/about-claude/models.
3. Argyris C, Schon DA. Organizational learning: A theory of action perspective. Reis: Revista Española de Investigaciones Sociológicas. 1997;77(78):345–8. https://doi.org/10.2307/40183951.
4. Bansal G, Wu T, Zhou J, Fok R, Nushi B, Kamar E, Ribeiro MT, Weld D. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. Proc 2021 CHI Conf Hum Factors Comput Syst. 2021. https://doi.org/10.1145/3411764.3445717.
5. Bender EM, Gebru T, McMillan-Major A, Shmitchell S On the dangers of stochastic parrots: Can language models be too big? Proc 2021 ACM Conf Fairness Acc Transpar. 2021. https://doi.org/10.1145/3442188.3445922.
6. Bhardwaj A, Faraj S. A quality of mercy is not trained: The imagined vs. the practiced in healthcare process-specialized AI development. arXiv. 2025;2510:21843. https://doi.org/10.48550/arXiv.2510.21843.
7. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, Bernstein MS, Bohg J, Bosselut A, Brunskill E, Brynjolfsson E, Buch S, Card D, Castellon R, Chatterji N, Chen A, Creel K, Davis JQ, Demszky D Liang, P. On the opportunities and risks of foundation models. arXiv. 2021;210807258. https://doi.org/10.48550/arXiv.2108.07258.

8. Cinelli M, Morales GDF, Galeazzi A, Quattrociocchi W, Starnini M. The echo chamber effect on social media. Proc Natl Acad Sci. 2021;118(9):e2023301118. https://doi.org/10.1073/pnas.2023301118.
9. Doshi AR, Hauser OP. Generative AI enhances individual creativity but reduces the collective diversity of novel content. Sci Adv. 2024;10(28):eadn5290. https://doi.org/10.1126/sciadv.adn5290.
10. Dzindolet MT, Peterson SA, Pomranky RA, Pierce LG, Beck HP. The role of trust in automation reliance. Int J Hum Comput Stud. 2003;58(6):697–718. https://doi.org/10.1016/S1071-5819(03)00038-7.
11. European Parliament & Council of the European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official J Eur Union. 2024;L Series:1–6. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689.
12. Funda V. A systematic review of algorithm auditing processes to assess bias and risks in AI systems. J Infrastructure Policy Dev. 2025;9(2):11489. https://doi.org/10.24294/jipd11489.
13. Goddard K, Roudsari A, Wyatt JC. Automation bias: A systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc. 2012;19(1):121–7. https://doi.org/10.1136/amiajnl-2011-000089.
14. Guu K, Lee K, Tung Z, Pasupat P, Chang M-W. REALM: Retrieval-augmented language model pre-training. Proc 37th Int Conf Mach Learn Proc Mach Learn Res. 2020;119:3929–38. https://proceedings.mlr.press/v119/guu20a.html.
15. International Organization for Standardization. ISO/IEC 42001:2023: Information technology — Artificial intelligence — Management system. 2023. https://www.iso.org/standard/81230.html.
16. Kahneman D. Thinking, fast and slow. Farrar, Straus and Giroux; 2011.
17. Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W-t, Rocktäschel T, Riedel S, Kiela D. Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv Neural Inf Process Syst. 2020;33:9459–74. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.
18. Lieder F, Griffiths TL, Huys QJM, Goodman ND. The anchoring bias reflects rational use of cognitive resources. Psychon Bull Rev. 2018;25(1):322–49. https://doi.org/10.3758/s13423-017-1286-8.
19. Meta AI. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. 2025. https://ai.meta.com/blog/llama-4-multimodal-intelligence/.
20. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, Spitzer E, Raji ID, Gebru T Model cards for model reporting. Proc Conf Fairness Acc Transpar. 2019. https://doi.org/10.1145/3287560.3287596.
21. National Institute of Standards and Technology. Artificial intelligence risk management framework (AI RMF 1.0). 2023. https://doi.org/10.6028/NIST.AI.100-1.
22. OpenAI. GPT-4.1 model documentation. 2025. https://platform.openai.com/docs/models/gpt-4.1.
23. Padmakumar V, He H. Does writing with language models reduce content diversity? In: Proceedings of the Twelfth International Conference on Learning Representations (ICLR 2024). arXiv. 2024;2309:05196. https://doi.org/10.48550/arXiv.2309.05196.
24. Parasuraman R, Riley V. Humans and automation: Use, misuse, disuse, abuse. Hum Factors. 1997;39(2):230–53. https://doi.org/10.1518/001872097778543886.
25. Power M. The audit society: Rituals of verification. Oxford University Press; 1997/2011. https://doi.org/10.1093/acprof:oso/9780198296034.001.0001.
26. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners [Technical report]. OpenAI. 2019. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.
27. Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, Smith-Loud J, Theron D, Barnes P. Closing the AI accountability gap. Proc 2020 Conf Fairness Account Transpar. 2020. https://doi.org/10.1145/3351095.3372873.
28. Star SL, Ruhleder K. Steps toward an ecology of infrastructure. Inform Syst Res. 1996;7(1):111–34. https://doi.org/10.1287/isre.7.1.111.
29. Strathern M. Improving ratings: Audit in the British university system. Eur Rev. 1997;5(3):305–21. https://doi.org/10.1017/S1062798700002660.
30. Tversky A, Kahneman D. Judgment under uncertainty: Heuristics and biases. Science. 1974;185(4157):1124–31. https://doi.org/10.1126/science.185.4157.1124.
31. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. Adv Neural Inf Process Syst. 2017;30:5998–6008.
32. Weidinger L, Mellor J, Rauh M, Griffin C, Uesato J, Huang P-S, Cheng M, Glaese M, Balle B, Kasirzadeh A, Kenton Z, Brown S, Hawkins W, Stepleton T, Biles C, Birhane A, Haas J, Rimell L, Hendricks LA, Isaac W, Gabriel I. Ethical and social risks of harm from language models. arXiv. 2021;2112:04359. https://doi.org/10.48550/arXiv.2112.04359.

Publisher's note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.