 ***DOI:** https://doi.org/10.63108/IJLMH.12783*

INTERNATIONAL JOURNAL OF LAW MANAGEMENT & HUMANITIES

[ISSN 2581-5369]

Volume 9 | Issue 5

2026

Hermeneutic Sovereignty under Generative AI: A Humanities Framework for Protecting Human Meaning-Making from Interpretive Foreclosure

KWAN HONG TAN*

ABSTRACT

Generative artificial intelligence increasingly mediates not only what institutions decide, but how institutions interpret people. Large language models draft case summaries, student feedback, performance reviews, clinical notes, policy briefs, creative text and administrative explanations. Existing AI governance frameworks appropriately emphasise accuracy, fairness, transparency, accountability and human oversight, yet these categories do not fully capture a distinct humanities problem: a fluent machine-generated interpretation can become institutionally authoritative before the person concerned has meaningfully articulated, contextualised or contested the account. This paper develops a person-centred framework of hermeneutic sovereignty for generative AI. Using an interdisciplinary conceptual methodology, it synthesises philosophical hermeneutics, epistemic injustice, narrative identity, human factors research, empirical studies of generative AI and contemporary governance frameworks. Hermeneutic sovereignty is defined as the situated and relational standing and capacity of persons and communities to participate meaningfully in constructing, contextualising, contesting, pluralising, revising and, where appropriate, refusing interpretations of their own experiences, identities, intentions, reasons and circumstances when those interpretations are mediated by AI. The paper identifies four mechanisms of AI-mediated interpretive foreclosure: interpretive pre-emption, hermeneutic compression, authority laundering and recursive fixation. It then proposes six dimensions for evaluating hermeneutic sovereignty and an Interpretive Foreclosure Test for institutional workflows. The analysis shows that human oversight is insufficient when the human merely approves a machine-framed account. Responsible adoption requires governance of the meaning-making process itself, including human-first elicitation, source-to-summary traceability, visible uncertainty, plural framing, contestability, correction propagation and time-bounded interpretive records. The paper concludes that trustworthy AI must preserve not only a human role in decisions, but a

* Author is a Lecturer at Singapore University of Social Sciences, Singapore. ORCID: https://orcid.org/0009-0003-9276-2829

 [ISSN 2581-5369]

meaningful human standing in the production of the interpretations on which decisions depend.

Keywords: *Generative artificial intelligence, Hermeneutic sovereignty, Interpretive foreclosure, Epistemic injustice, Human agency, Narrative identity, AI governance, Humanities*

I. INTRODUCTION

Generative artificial intelligence is rapidly becoming an interpretive layer between persons and institutions. Large language models do not merely retrieve stored facts or calculate predetermined outcomes. They synthesise, reformulate, classify, summarise and narrate. In practical settings, this means they can transform a complaint into an administrative synopsis, a set of employee observations into a performance narrative, a student's draft into a diagnostic account of ability, an interview transcript into a case note, or a collection of sources into an apparently coherent explanation. The productivity gains can be substantial. Experimental evidence shows that generative AI can reduce completion time and improve average quality in professional writing tasks (Noy & Zhang, 2023). In creative tasks, it can also increase individual performance while making outputs more similar at the collective level (Doshi & Hauser, 2024). These benefits help explain why generative systems are becoming embedded in knowledge work rather than remaining optional writing aids.

The humanities problem begins where assistance becomes interpretation. A generated summary is never simply a shorter version of an original. It selects salience, orders causes, establishes tone, chooses categories and suppresses some ambiguities so that others can be foregrounded. Even where an AI system has no consciousness, intention or lived understanding, its output can perform an interpretive function once a human or institution uses that output as a representation of another person's circumstances. The ethically significant question is therefore not whether the model "understands" in a phenomenological sense. It is whether its linguistic synthesis becomes socially operative as understanding.

Contemporary AI governance is well equipped to ask whether a system is accurate, robust, fair, secure and accountable. The NIST AI Risk Management Framework, for example, provides a practical structure for identifying and managing risks across the AI lifecycle, while its generative AI profile extends that orientation to risks associated with generative systems (Autio et al., 2024; Tabassi, 2023). Human-centred guidance in education likewise emphasises ethical validation, human agency and appropriate institutional capacity (Miao & Holmes, 2023). Ethical frameworks such as AI4People identify autonomy, justice, explicability and other

principles that should inform responsible AI (Floridi et al., 2018). These frameworks are necessary. Yet a gap remains between having a human in the loop and preserving a human place in the production of meaning.

A person may formally retain a right to appeal while confronting an AI-generated description that has already organised the institutional record. A manager may technically remain responsible for a performance review while beginning from a fluent AI narrative that frames an employee as disengaged. A teacher may retain final grading authority while an AI-generated diagnostic paragraph quietly establishes what counts as the student's weakness. Once such framings become the default representation, later human judgment can be constrained by the language, categories and causal story already supplied. The resulting risk is not only decision error. It is interpretive foreclosure, in which one provisional account hardens too early and narrows the space in which alternative meanings can be articulated and heard.

This paper asks: how should institutions govern generative AI when the system mediates interpretations of persons, experiences and reasons, rather than merely supplying information or automating routine tasks? A second question follows: what conditions must be preserved if affected persons are to remain meaningful participants in the interpretation of their own lives?

The paper develops a humanities framework of hermeneutic sovereignty to answer these questions. The term is not presented as a claim to lexical novelty. Recent scholarship has used related sovereignty language in different domains. Oziev (2026) develops hermeneutic sovereignty as a constitutional court's capacity to manage legal doubt and produce authoritative legal certainty. Sakabe (2026) uses interpretive sovereignty to describe human creative subjectivity in an analysis of artistic appropriation in the era of generative AI. These uses focus respectively on institutional legal authority and creative authorship. The present paper develops a different, person-centred and relational concept: the standing and capacity of affected persons and communities to participate in the construction, contestation and revision of AI-mediated interpretations that may shape institutional treatment.

The contribution is fourfold. The paper distinguishes hermeneutic sovereignty from adjacent forms of agency, theorises AI-mediated interpretive foreclosure through four mechanisms, proposes six diagnostic dimensions, and translates the framework into governance requirements through an Interpretive Foreclosure Test and lifecycle safeguards.

This development extends earlier work that framed algorithmic governance as a humanities problem of agency, dignity and social recognition (Tan, 2026a), and work on cognitive sovereignty that asks whether users retain the capacity to understand, verify, contest and

 [ISSN 2581-5369]

responsibly own AI-assisted knowledge claims (Tan, 2026b). The present argument narrows the lens to a different object: meaning itself. A person can understand an AI output and still be wronged by the institutional authority given to its interpretation. Conversely, a person can benefit from AI-assisted expression while retaining strong hermeneutic sovereignty if the system expands rather than closes interpretive possibilities. The normative goal is therefore neither AI abstinence nor human interpretive monopoly. It is to preserve a fair and revisable ecology of meaning in which machine synthesis remains answerable to human subjects.

II. METHODOLOGY

A. Interdisciplinary conceptual method

This is a normative and conceptual humanities study rather than an empirical validation study or a systematic literature review. Its method is interdisciplinary conceptual synthesis. The analysis brings together six bodies of material whose intersection is necessary to identify the problem: philosophical hermeneutics, narrative identity, epistemic injustice, human factors and cognitive offloading, empirical generative AI research, and contemporary AI governance guidance. The purpose is not to collapse these traditions into a single theory. It is to use them as mutually constraining lenses for a new institutional problem.

The hermeneutic tradition supplies an account of understanding as historically situated, dialogical and revisable rather than a purely technical act of extracting fixed meaning (Gadamer, 2004). Ricoeur's account of narrative identity is relevant because persons are not only objects described from outside; they participate in narrating continuity, action and responsibility across time (Ricoeur, 1992). Epistemic injustice scholarship shows that power affects who is heard as a knower and what interpretive resources are available for making sense of experience (Dotson, 2011; Fricker, 2007; Medina, 2013; Pohlhaus, 2012). Recent AI scholarship extends these concerns to algorithmic profiling and generative systems (Kay et al., 2024; Milano & Prunkl, 2025).

Human factors and cognitive offloading research provide a complementary mechanism-level perspective. Reliance on automation depends on trust calibration, system characteristics and context (Lee & See, 2004). Cognitive offloading can reduce immediate mental demands while also redistributing the work required for memory, evaluation and control (Risko & Gilbert, 2016). These insights matter because interpretive authority can shift without explicit delegation. A user may remain nominally responsible while the first coherent wording supplied by an AI system structures subsequent judgment.

 [ISSN 2581-5369]

Empirical studies of generative AI provide evidence about benefits and homogenising pressures. Generative AI can increase productivity and apparent quality in professional tasks (Noy & Zhang, 2023), and it can increase individual creative performance while reducing collective diversity (Doshi & Hauser, 2024). In science, Messeri and Crockett (2024) warn that AI tools can produce illusions of understanding and scientific monocultures, where apparent productivity coexists with reduced epistemic diversity. These findings do not prove interpretive foreclosure, but they support the plausibility of mechanisms through which fluent assistance can narrow human exploration while improving surface-level performance.

B. Scope, evidential discipline and limitations

The paper focuses on text-generating systems used in institutional or quasi-institutional settings where outputs may affect how persons are represented, evaluated, taught, managed, served or governed. Examples include education, employment, public administration, professional knowledge work and creative industries. The argument also has relevance to legal and healthcare settings, but it does not attempt domain-specific doctrinal or clinical analysis. Such applications require additional legal, professional and empirical work.

The conceptual method follows three disciplines. Claims about existing theory are separated from the paper's new constructs. Empirical studies are used only for the phenomena they establish, so productivity gains, output similarity, automation reliance and illusions of understanding are evidence of relevant tendencies rather than direct proof of institutional injustice. Finally, the framework is designed for empirical challenge: its dimensions generate observable questions about workflow design, record persistence, contestation and revision.

The principal limitation is therefore intentional: the paper offers a theory and governance framework, not causal estimates. Future work should operationalise its dimensions through experiments, document audits, interviews and longitudinal studies. This limitation is preferable to giving numerical precision to a concept that has not yet been empirically calibrated.

III. Interpretation, agency and epistemic power

A. Understanding is not extraction

A foundational insight of philosophical hermeneutics is that interpretation cannot be reduced to extracting a neutral meaning that already exists in fully determinate form. Understanding occurs from within historically and linguistically situated horizons. Gadamer (2004) emphasises that interpretation is shaped by prior understandings and becomes productive through dialogue. This

does not imply that every interpretation is equally valid. It means that the conditions under which meaning appears are themselves part of what must be examined.

Generative AI intensifies the institutional significance of this point because it can manufacture linguistic coherence at very low marginal cost. A case file containing uncertainty, contradiction and incomplete testimony can be transformed into a smooth paragraph. Smoothness can be useful, but it can also obscure the fact that a synthesis is an achievement of framing rather than a transparent view onto reality. The more persuasive the prose, the easier it becomes to mistake an interpretation for the underlying material itself.

Bender et al. (2021) caution against attributing human-like understanding to language models whose outputs arise from statistical patterning over linguistic data. For the present argument, however, the absence of model understanding does not eliminate the hermeneutic problem. It relocates it. If a model cannot bear responsibility for understanding, then the human and institutional processes that adopt its representations become more important. The central issue becomes how generated language acquires authority in social practices.

B. Narrative identity and the right to remain revisable

Ricoeur's hermeneutics of the self provides a second foundation. Personal identity is not exhausted by static attributes. It is narrated through action, memory, character, promise and revision across time (Ricoeur, 1992). Institutional descriptions participate in this narrative environment. Labels such as "high risk," "unmotivated," "non-compliant," "gifted," "unreliable," or "leadership potential" may be administratively convenient, but they can also become durable narrative positions from which later conduct is interpreted.

Generative AI raises the stakes because narrative production can become automated and recursive. Once an AI-generated summary enters a record, subsequent systems may retrieve it, compress it again, use it as context for another generation, or present it to future decision-makers. A provisional narrative can therefore acquire temporal durability without anyone consciously deciding that it should. This is a distinct form of power: not the power to make one decision, but the power to establish the story from which later decisions begin.

Hermeneutic sovereignty responds to this temporal dimension by insisting on revisability. Persons cannot reasonably demand control over every external interpretation of them. Social life necessarily involves interpretation by others. But when an institution operationalises an interpretation in ways that affect opportunities, obligations or recognition, the affected person should retain meaningful standing to correct, contextualise and contest it. This standing is especially important where the record will persist.

 [ISSN 2581-5369]

C. Epistemic injustice and interpretive asymmetry

Fricker (2007) distinguishes testimonial injustice, where prejudice causes a speaker to receive an unfair credibility deficit, from hermeneutical injustice, where unequal participation in collective meaning-making leaves some experiences inadequately intelligible. Subsequent scholarship has deepened the relational and structural dimensions of these harms. Dotson (2011) analyses forms of silencing that occur when audiences fail to meet speakers' epistemic vulnerabilities. Pohlhaus (2012) shows how dominant knowers can refuse interpretive resources developed from marginalised experience. Medina (2013) emphasises epistemic resistance and the importance of friction among differently situated perspectives.

Algorithmic systems can interact with these structures rather than merely reproducing biased outputs. Milano and Prunkl (2025) argue that algorithmic profiling can generate hermeneutical injustice through epistemic fragmentation, making it harder for individuals to identify and conceptualise shared harms. Kay et al. (2024) develop the idea of generative algorithmic epistemic injustice, including hermeneutical ignorance and access injustice within AI-mediated knowledge ecosystems. These accounts demonstrate that the relevant harm is not always an incorrect proposition. It can lie in the organisation of interpretive resources and the conditions under which people can make sense of their situations.

The present framework adds a related but distinct concern. Generative systems do not only fragment epistemic environments. They can also pre-assemble them. An institution may receive a coherent AI-generated interpretation before it hears the affected person's own formulation. This creates an asymmetry of sequence and format. The machine narrative arrives early, polished and aligned with institutional categories; the person's reply arrives later, perhaps emotionally, incompletely or outside the preferred template. Equal formal access to the process does not neutralise this asymmetry.

D. From human agency and cognitive sovereignty to hermeneutic sovereignty

Tan (2026a) argues that algorithmic governance should be assessed by whether affected persons retain interpretive agency, contestatory power, relational recognition and meaningful participation. Tan (2026b) develops cognitive sovereignty as the capacity to originate, evaluate, contest, retain and responsibly own knowledge claims even when computational assistance is used. Both approaches point toward a broader question of human control under AI mediation.

Hermeneutic sovereignty narrows and deepens one part of that problem. It is not primarily about whether a person can verify a factual claim, nor whether a human remains formally able to act. It concerns who has standing in determining what an experience means, which context is

 [ISSN 2581-5369]

relevant, which categories apply, and whether a dominant interpretation can be reopened. It therefore occupies the space between epistemic agency and social recognition.

This paper defines hermeneutic sovereignty as the situated and relational standing and capacity of persons and communities to participate meaningfully in constructing, contextualising, contesting, pluralising, revising and, where appropriate, refusing interpretations of their own experiences, identities, intentions, reasons and circumstances when those interpretations are mediated by generative AI.

Three qualifications are essential. First, sovereignty is situated, not absolute. Its requirements rise with stakes, persistence, asymmetry of power and difficulty of reversal. Second, it is relational, not atomistic. Meaning is socially formed, and institutions legitimately interpret conduct for shared purposes. Third, it concerns standing as well as capacity. A person may have the intellectual ability to contest an interpretation but lack an authorised channel through which the contest can matter. Conversely, an appeal channel may formally exist while the person lacks access to the source material or sufficient explanation to use it.

IV. Generative AI as interpretive infrastructure

A. From tools to infrastructures of meaning

A spelling checker changes text without normally determining what the author means. A large language model can occupy a much more consequential role. It can propose the issue, select the facts, generate the causal link, assign the tone and recommend the conclusion. When such outputs are embedded in routine workflows, the system becomes interpretive infrastructure: an environment through which meaning is produced and circulated.

This infrastructural role is easy to overlook because the interface often presents generation as assistance. The human clicks “summarise,” “draft,” “analyse,” or “improve.” Yet each command can shift the cognitive and hermeneutic starting point. Research on automation shows that trust and reliance are shaped by context, perceived competence and system presentation (Lee & See, 2004). Cognitive offloading research likewise shows that people strategically externalise cognitive work, sometimes based on imperfect metacognitive judgments (Risko & Gilbert, 2016). In a generative setting, what is offloaded can include not only memory or calculation but framing.

The shift matters because first framings can anchor later deliberation. Even when a human edits the output, the generated structure determines what must be actively noticed and overturned. A person reviewing a fluent summary is not in the same epistemic position as one constructing an

account from primary material. The former task is correction; the latter is interpretation. What the generated text omits may never become salient enough to challenge.

B. Fluency, compression and the social appearance of authority

Generative systems produce language with a degree of fluency that can exceed the certainty of the underlying evidence. This is not merely a hallucination problem. A fully factually accurate summary may still be normatively incomplete. It may omit hesitation, disagreement, context or alternative causal explanations in order to achieve coherence.

Messeri and Crockett (2024) show how AI-supported science can encourage illusions of understanding, where researchers feel that they comprehend more than they actually do, while the scientific community becomes vulnerable to monocultures. A parallel risk appears in institutional interpretation. The danger is not simply that a generated account is false. It is that an account becomes cognitively satisfying enough to reduce the perceived need for further interpretation.

Empirical findings on creativity illustrate another relevant pattern. Doshi and Hauser (2024) find that AI assistance can improve individual outputs while reducing collective diversity. Institutional interpretation can exhibit an analogous tension. Standardised generated language may raise average clarity and consistency, yet also reduce the variety of ways in which ambiguous situations are described. When a common model, prompt library or organisational template structures many cases, the institution may become more linguistically consistent while becoming less hermeneutically plural.

C. Meaning is governed through workflow

The governance problem therefore cannot be solved by output disclaimers alone. A label stating that "AI may make mistakes" does little if the workflow gives machine-generated text priority, persistence and administrative convenience. Interpretive power is governed through sequence, interface defaults, record architecture, escalation rules and the distribution of labour.

Consider two superficially similar workflows. In the first, an employee writes a self-assessment, a manager records observations, the two discuss disagreements, and AI is then used to improve clarity while preserving provenance. In the second, an AI system synthesises metrics and messages into a draft performance narrative before either party articulates an account. The manager edits the draft and the employee can comment after the rating is proposed. Both workflows retain a "human in the loop." Only the first gives human accounts procedural priority. This distinction motivates the concept of interpretive foreclosure.

 [ISSN 2581-5369]

V. AI-MEDIATED INTERPRETIVE FORECLOSURE

AI-mediated interpretive foreclosure occurs when a provisional machine-generated interpretation becomes sufficiently authoritative, early, compressed or persistent that it narrows the practical space for alternative meanings before affected persons can meaningfully articulate, contest or revise them. It is not identical to hermeneutical injustice. Foreclosure names a workflow mechanism; it becomes an injustice when that mechanism unfairly diminishes a person's or group's interpretive standing, often under conditions of institutional power. Foreclosure need not be deliberate and can emerge from convenience, incentives and the rhetorical force of fluent synthesis. Four mechanisms are especially important.

A. Interpretive pre-emption

Interpretive pre-emption occurs when AI supplies the first organised account of a person or event before the affected person, responsible professional or relevant community has produced an independent interpretation. Sequence matters because initial frames structure attention. Once a narrative identifies the "main issue," later information tends to be read as confirming, qualifying or disputing that issue rather than establishing a different one.

Pre-emption is most problematic in high-stakes contexts involving identity, intention, responsibility or vulnerability. A generated disciplinary summary may frame conduct as defiance rather than confusion. A generated student diagnostic may frame poor performance as lack of understanding rather than language difficulty or unfamiliarity with the assessment form. The alternative interpretation need not be correct. The institutional concern is that the privilege of first meaning should not be silently allocated to the machine.

B. Hermeneutic compression

Hermeneutic compression is the reduction of ambiguous, context-rich or plural experience into a compact representation optimised for administrative use. All institutions compress information. The distinctive risk of generative AI is the speed, scale and linguistic completeness with which compression can occur.

Compression becomes harmful when it removes precisely the ambiguity that should remain visible. Erickson and Gregory (2025) make a related jurisprudential argument that algorithmic demands for specification can undermine the productive vagueness of legal concepts and transfer interpretive authority to extra-legal processes. Their focus is legal language and the open texture of law. The present framework extends the concern to representations of persons.

 [ISSN 2581-5369]

Some human situations should not be rendered fully determinate merely because an information system performs better with definite categories.

The right safeguard is not maximal detail. Overdocumentation can itself burden and surveil. The normative requirement is contextual integrity: the representation should preserve the distinctions, uncertainty and dissent that are material to fair interpretation.

C. Authority laundering

Authority laundering occurs when human normative judgment is presented through machine-generated prose in a way that makes the judgment appear more neutral, objective or technically compelled than it is. The AI does not possess institutional legitimacy, yet its fluency can mask the point at which discretionary choices entered the process.

This can happen in mundane ways. A manager may ask AI to "objectively explain" why an employee is not ready for promotion and use the generated rationale as though it were discovered rather than composed. An administrator may classify a complaint through AI-generated categories and later treat those categories as natural features of the case. A teacher may use AI feedback as evidence of weakness rather than as one interpretation of the student's work.

Authority laundering weakens accountability because it obscures authorship. The human can attribute the wording to the system; the system cannot answer for the judgment. Responsible governance must therefore preserve traceable human ownership of normatively significant interpretations.

D. Recursive fixation

Recursive fixation occurs when a generated interpretation persists across records and later becomes input for additional interpretations, gradually acquiring authority through repetition. A first summary may be copied into a second report, retrieved into a new model context, paraphrased in correspondence, and reintroduced as "background." Each repetition can make the original framing harder to distinguish from independently established fact.

This mechanism is particularly important because correction is often local while propagation is systemic. A person may successfully amend one record, yet an earlier AI-generated description may remain in downstream databases, cached summaries or derivative documents. If later systems treat the repeated statement as corroboration, a single initial interpretation can become self-confirming.

 [ISSN 2581-5369]

Recursive fixation transforms a temporal convenience into a governance problem. It also connects hermeneutic sovereignty to data lifecycle management. The relevant question is not only who can correct a record, but whether the correction follows the interpretation wherever it has travelled.

VI. THE HERMENEUTIC SOVEREIGNTY FRAMEWORK

A. Six dimensions

Hermeneutic sovereignty can be evaluated through six dimensions. They are not intended as a psychometric scale. They function as normative and diagnostic dimensions that can be translated into domain-specific indicators.

Dimension	Core question	Typical failure	Governance implication
Narrative authorship	Did affected persons have a meaningful opportunity to state or frame their own account before consequential AI synthesis?	Machine framing becomes the default starting narrative.	Use human-first elicitation where identity, intention or responsibility is at stake.
Contextual integrity	Does the AI-mediated representation preserve materially relevant context, ambiguity and disagreement?	Rich circumstances are compressed into administratively convenient categories.	Require source-to-summary traceability and visible uncertainty.
Interpretive contestability	Can affected persons challenge not only a decision, but also the interpretation on which it relies?	Appeals address outcomes while the underlying narrative remains fixed.	Provide annotation, correction and counter-narrative channels.
Plurality preservation	Can multiple reasonable interpretations remain visible when evidence is ambiguous?	One fluent synthesis suppresses alternatives too early.	Require alternative framings or dissent markers in high-ambiguity cases.
Temporal revisability	Can interpretations change as new evidence, reflection or circumstances emerge?	Early labels persist and become self-confirming.	Use review dates, expiry rules and correction propagation.
Relational accountability	Is a responsible human or institution answerable for adopting and acting on the interpretation?	Responsibility is diffused between user, model, vendor and workflow.	Name accountable adopters and record reasons for consequential uptake.

Table 1: Dimensions of hermeneutic sovereignty

These dimensions jointly clarify why a simple "human review" requirement is too weak. A human reviewer may retain formal authority while narrative authorship, contextual integrity and plurality have already been lost. Conversely, a workflow can use substantial AI assistance while preserving sovereignty if human accounts are elicited first, uncertainty is maintained,

sources remain inspectable, alternatives can be expressed, corrections propagate and accountable humans own the final interpretation.

B. Standing, capacity and institutional uptake

The framework treats sovereignty as the interaction of three elements: standing, capacity and uptake. Standing concerns whether a person is recognised as entitled to contribute to the interpretation. Capacity concerns whether the person has practical resources to do so, including access to records, sufficient explanation, language support, time and cognitive opportunity. Uptake concerns whether the institution is obliged to consider the contribution in a way that can alter the operative account.

All three are necessary. A comment box offers little uptake if no reviewer must address the correction. Explanation does not create standing if the affected person cannot append a counter-account. A formal right to contest can also be hollow if sources are unavailable or the generated narrative has already propagated.

This triadic structure also prevents an overly individualistic account. Hermeneutic sovereignty may be exercised collectively, particularly where groups need shared interpretive resources to identify structural harms. The epistemic injustice literature shows why collective sense-making matters for experiences that are difficult to name within dominant categories (Fricker, 2007; Medina, 2013; Milano & Prunkl, 2025). Institutions should therefore be attentive not only to individual correction mechanisms but also to whether repeated cases can be compared and contested collectively.

C. The interpretive foreclosure test

For practical governance, the framework proposes an Interpretive Foreclosure Test. Before deploying generative AI in a workflow that produces interpretations of persons, institutions should ask four questions:

1. Priority: does the machine-generated frame arrive before a relevant human account has been independently elicited?
2. Compression: does the generated representation remove material ambiguity, disagreement or context in order to fit a decision process?
3. Authority: is the generated text likely to be treated as neutral evidence rather than as a contestable synthesis adopted by an accountable human?
4. Persistence: can the interpretation propagate or influence later decisions without a reliable mechanism for revision and correction propagation?

 [ISSN 2581-5369]

A positive answer to one question is not automatically unacceptable. The test is not a prohibition rule. It is a trigger for stronger safeguards. Risk becomes especially serious when all four conditions align: the machine speaks first, compresses complexity, acquires institutional authority and persists over time. That configuration creates a pathway from convenience to interpretive domination.

The test complements rather than replaces existing AI risk management. The NIST risk-management approach is lifecycle-oriented and deliberately broad (Autio et al., 2024; Tabassi, 2023). The Interpretive Foreclosure Test adds a humanities-specific lens for workflows in which the object at risk is the fairness and revisability of meaning-making itself.

VII. Applications

A. Education

Education provides a clear illustration because generative AI can function simultaneously as tutor, evaluator, editor and diagnostic narrator. UNESCO guidance correctly emphasises human-centred use, capacity development and pedagogical validation (Miao & Holmes, 2023). Hermeneutic sovereignty adds a specific requirement: students should not become passive objects of AI-generated interpretations of their ability, motivation or learning difficulty.

A high-sovereignty workflow might begin with student self-explanation, evidence of work and teacher observation. AI could then propose alternative feedback formulations, identify patterns or raise questions. The teacher would remain responsible for interpretation and the student would be able to respond to the account. A low-sovereignty workflow would generate a diagnostic profile from submissions and behavioural data, present that profile as a coherent description, and ask the student merely to accept an intervention.

The distinction is educationally substantive. AI literacy should include the capacity to challenge interpretations, not only verify factual outputs. This extends cognitive sovereignty from knowing and checking toward self-interpretation in institutional contexts (Tan, 2026b).

B. Work, performance and organisational identity

Generative AI is particularly attractive in performance management because managers face large documentation burdens. Yet performance reviews are not neutral compilations of facts. They narrate contribution, potential, attitude and trajectory. When AI synthesises emails, project data or prior feedback, it can transform scattered observations into an authoritative organisational identity.

 [ISSN 2581-5369]

Hermeneutic sovereignty does not require employees to control performance judgments. Organisations have legitimate evaluative authority. It requires that machine-assisted narratives remain attributable, contestable and revisable. Employees should know when material descriptions are AI-generated or AI-synthesised, have access to the relevant source basis, and be able to append a substantive response that remains linked to the record. Managers should record why they adopted a consequential interpretation rather than relying on the authority of polished machine prose.

The framework also cautions against recursive fixation. A phrase such as "resistant to change" can migrate from one AI-assisted review into succession planning, development recommendations and later prompts. Without a propagation-aware correction process, an initially weak interpretation can become organisational memory.

C. Public administration and citizen narratives

Public administration routinely converts lived circumstances into forms, categories and eligibility criteria. Generative AI may improve accessibility by helping citizens explain complex situations, translating between registers and assisting caseworkers with document synthesis. These are significant benefits. In some cases, AI can enhance hermeneutic sovereignty by giving people language for experiences they struggled to articulate.

The same technology can pre-empt citizen narratives. If automated intake converts a free-text account into categories before a caseworker or citizen can review the transformation, administrative convenience may determine meaning. The risk is heightened by language barriers, disability, low institutional literacy or unequal power. Formal opportunities to speak do not guarantee fair interpretive uptake (Dotson, 2011; Fricker, 2007).

A sovereignty-preserving system should therefore display the transformation from source account to administrative summary, allow citizens or representatives to correct material changes, and preserve disputed interpretations rather than forcing a single settled narrative where the evidence remains contested. Where an interpretation has downstream consequences, correction should propagate across dependent records.

D. Creative and professional knowledge work

Creative work shows why hermeneutic sovereignty should not be understood as opposition to AI assistance. Sakabe's (2026) account of interpretive sovereignty in creative authorship and Doshi and Hauser's (2024) empirical findings both point toward a more nuanced relationship. Generative AI can supply prompts, alternatives and stylistic possibilities that expand an

individual's expressive range. The concern arises when the system becomes the default source of interpretation rather than an interlocutor within a human-led process.

For writers, researchers, consultants and analysts, one safeguard is deliberate plural prompting: use AI to generate competing frames, objections and counter-interpretations rather than one authoritative synthesis. Another is temporal separation: formulate an initial thesis or account before consulting a model, especially when originality or judgment is central. These practices introduce productive friction. They preserve the user's role as an originator and evaluator rather than turning review into acceptance of the first fluent proposal.

The same logic responds to the risk of monoculture identified by Messeri and Crockett (2024). Organisational knowledge systems should reward interpretive diversity, not merely faster convergence. In ambiguous domains, disagreement can be information.

VIII. Objections and boundary conditions

A. AI can expand, not only narrow, interpretive agency

The strongest objection is that generative AI can help people articulate meanings that institutions previously ignored. A person with limited confidence, language proficiency or specialist vocabulary may use AI to draft a complaint, organise memories or explore alternative framings. Generative AI can therefore increase access to interpretive resources and counter some forms of epistemic disadvantage.

This objection is correct and is incorporated into the framework. Hermeneutic sovereignty is not a preservationist demand that all meaning originate unaided in the individual. Human interpretation has always depended on language, cultural resources, institutions and other people. The normative distinction is between augmentation and foreclosure. AI supports sovereignty when it expands the person's repertoire while keeping alternatives visible and preserving the person's capacity to revise or refuse. It undermines sovereignty when the generated frame becomes difficult to escape because of sequence, authority or persistence.

B. Sovereignty may sound too individualistic

A second objection concerns the language of sovereignty. Meaning is relational, and no individual has exclusive ownership over how others interpret conduct. An employee cannot unilaterally define whether performance met organisational expectations; a citizen cannot dictate the legal meaning of conduct; a student cannot decide the academic standard by self-description.

 [ISSN 2581-5369]

The framework uses sovereignty in a deliberately limited sense: protected standing within interpretive relations, not unilateral semantic control. It supports claims to speak before definitive framing, know the basis of consequential interpretations, contest material misdescription, preserve warranted ambiguity and obtain revision when the operative record is no longer justified. These claims coexist with legitimate external judgment.

C. Human interpretation has always been biased

A third objection is that interpretive foreclosure predates AI. Bureaucracies, teachers, managers and courts have always reduced complexity into categories. Human decision-makers also anchor, stereotype and repeat past narratives. Why treat generative AI as a distinct problem?

AI does not create interpretation or injustice for the first time. It changes their scale, speed, standardisation and persistence. Generative systems can produce polished summaries instantly, propagate them across organisational systems and normalise a shared style of reasoning, while responsibility becomes easier to diffuse. The framework responds to these changed conditions of interpretive power rather than assuming earlier institutions were hermeneutically just.

D. Not every summary requires a sovereignty protocol

A final objection is practical. Most organisations use AI for low-stakes summarisation, scheduling, drafting and search. Requiring elaborate contestation for every generated sentence would destroy the productivity benefits that motivate adoption.

The framework is risk-sensitive. Strong safeguards are warranted when interpretations are person-directed, consequential, persistent, difficult to reverse, or produced under significant power asymmetry. A summary of a public meeting agenda does not require the same controls as an AI-generated explanation for disciplinary action. The Interpretive Foreclosure Test helps identify where the conditions for foreclosure are present. Governance should concentrate on those points rather than impose uniform friction everywhere.

IX. Governance implications: hermeneutic sovereignty by design

The central governance implication is that institutions should manage the meaning chain, not only the model. The relevant lifecycle begins with how human experience is elicited, continues through generation and human adoption, and extends into records, appeals, reuse and later system prompts. A technically safe model can still participate in an unfair meaning chain if the surrounding workflow gives its interpretations premature or irreversible authority.

 [ISSN 2581-5369]

Stage	Primary risk	Safeguard	Evidence for audit
Elicitation	Interpretive pre-emption	Obtain human-first accounts in high-stakes narrative contexts.	Timestamped source narratives; record of who framed the issue first.
Generation	Hermeneutic compression	Preserve uncertainty, dissent and material context; generate alternatives where appropriate.	Source-to-summary trace; uncertainty markers; alternate framings.
Adoption	Authority laundering	Require an accountable human to state reasons for consequential adoption.	Named adopter; documented rationale; provenance of AI contribution.
Communication	Apparent neutrality	Disclose material AI synthesis and distinguish evidence from interpretation.	User-facing provenance statement; accessible source basis.
Contestation	Outcome-only appeal	Allow challenge to the interpretation itself and attach counter-narratives.	Correction logs; response times; disposition of contested framings.
Persistence	Recursive fixation	Apply review dates, expiry rules and correction propagation across dependent records.	Version history; downstream correction confirmation; stale-record alerts.
Audit	Invisible loss of plurality	Review recurrent categories, disagreement suppression and correction recurrence.	Rates of contested summaries, repeated labels, unresolved divergence and reappearance after correction.

Table 2: Hermeneutic sovereignty safeguards across the meaning chain

Several design principles follow. First, human-first elicitation should be the default where systems interpret intention, identity, responsibility, vulnerability or disputed experience. This does not prohibit AI-assisted expression. A person may use AI to organise an account, but the system should not silently establish the institutional frame before the person has a chance to speak.

Second, institutions need source-to-summary traceability. Reviewers and affected persons should be able to see which underlying materials support major interpretive claims and which elements were introduced through synthesis. This is different from conventional explainability, which often focuses on model logic. The relevant transparency concerns representational transformation.

Third, systems should preserve visible uncertainty and plurality. Where evidence supports multiple reasonable interpretations, interfaces should resist collapsing them into a single

confident paragraph. Alternative framings, unresolved questions and dissent markers can be treated as governance features rather than defects.

Fourth, institutions should establish interpretive contestability. Appeals should not be limited to final outcomes. A person should be able to challenge an intermediate characterisation that will continue to influence later decisions. Where a counter-account cannot replace the institution's view, it should at least remain linked to the operative record in a way visible to future reviewers.

Fifth, organisations require propagation-aware correction. Revision must follow derived summaries, dashboards and downstream systems. Otherwise, the formal right to correction is defeated by recursive fixation.

Sixth, accountable human adoption must be substantive. The human reviewer should not merely click approval. For consequential interpretations, the institution should know who adopted the account, what evidence was relied upon, what material uncertainty remained, and whether the affected person's account was considered. This re-establishes responsibility at the point where machine prose becomes institutional action.

These requirements also clarify the limits of general AI ethics principles. Autonomy, justice and explicability remain essential (Floridi et al., 2018), and risk-management frameworks provide valuable organisational scaffolding (Autio et al., 2024; Tabassi, 2023). Hermeneutic sovereignty does not compete with them. It specifies a domain of human agency that can otherwise disappear inside broad principles. A system can be documented, privacy-preserving and technically reliable while still narrowing the interpretive standing of those it represents.

X. Research agenda

The framework generates several empirical questions. Experiments can test sequence effects by comparing human-first, AI-first and parallel workflows, measuring decision accuracy alongside interpretive diversity, correction rates, confidence calibration and the ability to generate alternatives after exposure to an AI frame.

Second, field studies can examine compression loss. Researchers could compare original narratives with AI-generated summaries and code which contextual distinctions, uncertainties and dissenting elements disappear. Particular attention should be paid to cases involving culturally specific language or experiences that do not map neatly onto institutional categories.

Third, organisational audits can study recursive fixation by tracing how descriptions propagate across records and whether corrections reliably travel to downstream systems. This would turn a conceptual claim into a measurable data-governance problem.

 [ISSN 2581-5369]

Fourth, research should distinguish productive interpretive friction from unnecessary administrative burden. Some friction is valuable because it forces source inspection, comparison and reason-giving. Too much friction may reduce access, delay services or encourage informal workarounds. The design challenge is to place effort at epistemically diagnostic points rather than merely add steps.

Fifth, future studies should test whether the six dimensions of hermeneutic sovereignty form a coherent construct across domains or require sector-specific weighting. Education may place particular emphasis on narrative authorship and developmental revisability; public administration may prioritise contestability and correction propagation; creative work may emphasise plurality preservation; employment may require strong relational accountability.

Comparative humanities research should also examine how hermeneutic sovereignty changes across cultures and institutional traditions. Concepts of personhood, authority, narrative responsibility and collective identity vary. The framework's relational formulation is a starting point for such comparison, not a universal model of the autonomous individual.

XI. Conclusion

Generative AI changes the governance of interpretation. Its significance lies not only in faster drafting, better search or automated recommendation, but in its growing ability to supply the language through which institutions understand people. When machine-generated interpretations become the starting point for evaluation, administration or professional judgment, a new humanities problem emerges: the person may remain formally present while losing practical standing in the construction of meaning.

This paper has developed hermeneutic sovereignty as a person-centred and relational framework for that problem. Hermeneutic sovereignty does not mean that individuals control every interpretation of themselves, nor that AI should be excluded from meaning-making. It means that persons and communities should retain meaningful standing and capacity to construct, contextualise, contest, pluralise, revise and, where appropriate, refuse AI-mediated interpretations that shape institutional treatment.

Four mechanisms explain how this standing can be weakened: interpretive pre-emption, hermeneutic compression, authority laundering and recursive fixation. Together they describe AI-mediated interpretive foreclosure, a process through which provisional machine framings can become authoritative before adequate human interpretation has occurred. The six dimensions of narrative authorship, contextual integrity, interpretive contestability, plurality preservation, temporal revisability and relational accountability translate the theory into

 [ISSN 2581-5369]

institutional questions. The Interpretive Foreclosure Test further identifies when workflow design should trigger stronger safeguards.

The resulting governance principle is simple but demanding: human oversight is not enough when the human only oversees a meaning already framed by the machine. Trustworthy AI requires governance of the meaning chain. Institutions should preserve human-first elicitation where stakes are high, maintain source-to-summary traceability, keep uncertainty and alternatives visible, enable interpretive contestation, propagate corrections and locate final responsibility in accountable humans.

The broader humanities stakes are considerable. A society does not preserve human agency merely by reserving the final click for a person. It preserves agency by ensuring that people can still participate in the stories, categories and reasons through which institutions recognise them. As generative AI becomes an ordinary language of administration and knowledge work, protecting that interpretive standing will be as important as protecting accuracy, privacy or procedural review. The future of human-centred AI depends not only on who decides, but also on who gets to help determine what the situation means before a decision is made.

XII. References

Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). *Artificial intelligence risk management framework: Generative artificial intelligence profile* (NIST AI 600-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.600-1

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Doshi, A. R., & Hauser, O. P. (2024). Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances, 10*(28), Article eadn5290. https://doi.org/10.1126/sciadv.adn5290

Dotson, K. (2011). Tracking epistemic violence, tracking practices of silencing. *Hypatia, 26*(2), 236–257. https://doi.org/10.1111/j.1527-2001.2011.01177.x

Erickson, J., & Gregory, M. (2025). Against algorithmic clarity: Law beyond specification. *International Journal for the Semiotics of Law*. Advance online publication. https://doi.org/10.1007/s11196-025-10379-5

 [ISSN 2581-5369]

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People: An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198237907.001.0001

Gadamer, H.-G. (2004). *Truth and method* (J. Weinsheimer & D. G. Marshall, Trans.; 2nd rev. ed.). Continuum.

Kay, J., Kasirzadeh, A., & Mohamed, S. (2024). Epistemic injustice in generative AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 7*(1), 684–697. https://doi.org/10.1609/aies.v7i1.31671

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Medina, J. (2013). *The epistemology of resistance: Gender and racial oppression, epistemic injustice, and the social imagination*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199929023.001.0001

Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature, 627*(8002), 49–58. https://doi.org/10.1038/s41586-024-07146-0

Miao, F., & Holmes, W. (2023). *Guidance for generative AI in education and research*. UNESCO. https://doi.org/10.54675/EWZM9535

Milano, S., & Prunkl, C. (2025). Algorithmic profiling as a source of hermeneutical injustice. *Philosophical Studies, 182*(1), 185–203. https://doi.org/10.1007/s11098-023-02095-2

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Science, 381*(6654), 187–192. https://doi.org/10.1126/science.adh2586

Oziev, T. T. (2026). Towards a methodology of analyzing defensive constitutionalism: The model of the branched hermeneutic cycle and the concept of hermeneutic sovereignty. *Izvestiya of Saratov University. Economics. Management. Law, 26*(2), 206–216. https://doi.org/10.18500/1994-2540-2026-26-2-206-216

 [ISSN 2581-5369]

Pohlhaus, G., Jr. (2012). Relational knowing and epistemic injustice: Toward a theory of willful hermeneutical ignorance. *Hypatia, 27*(4), 715–735. https://doi.org/10.1111/j.1527-2001.2011.01222.x

Ricoeur, P. (1992). *Oneself as another* (K. Blamey, Trans.). University of Chicago Press. https://doi.org/10.7208/chicago/9780226844541.001.0001

Risko, E. F., & Gilbert, S. J. (2016). Cognitive offloading. *Trends in Cognitive Sciences, 20*(9), 676–688. https://doi.org/10.1016/j.tics.2016.07.002

Sakabe, H. (2026). A study on creative mechanisms based on case analysis of Suzy Lee's appropriation: Focusing on interpretive sovereignty in the era of generative AI. *Design Convergence Study, 25*(2), 35–50.

Tabassi, E. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

Tan, K. H. (2026a). Human agency under algorithmic governance: A humanities framework for law, management and public life. *International Journal of Law Management & Humanities, 9*(4), 559–571.

Tan, K. H. (2026b). *The cognitive sovereignty threshold: An interdisciplinary humanities and sciences framework for human agency in generative AI-assisted knowledge work* [Preprint]. ResearchGate. https://doi.org/10.13140/RG.2.2.21311.27045

*****

 [ISSN 2581-5369]