https://doi.org/10.37708/ezs.swu.bg.v24i3.3

AI DOES NOT JUDGE: EPISTEMIC ARCHAEOLOGY AND GOVERNANCE OF MEANING IN THE HUMANITIES AND SOCIAL SCIENCES

Federico Irace

University of Pavia, Italy
Email: federico.irace01@universitadipavia.it

ORCID ID: 0009-0001-0499-9590

ABSTRACT: The adoption of GenAI in research is reshaping the epistemic conditions of knowledge production. While LLMs are often perceived as neutral agents, this paper argues that the humanities and social sciences (HSS) serve as critical infrastructures for governing AI-mediated inquiry. Drawing on sociotechnical pragmatism, the article introduces epistemic archaeology: an interpretive orientation that treats AI outputs as stratified artifacts shaped by cultural and statistical histories. Rather than seeking validation in convergence, this approach interprets model divergence as signals of training data and epistemic constraints. By foregrounding interpretation, contextualization, and epistemic literacy, the work establishes the foundational coordinates for AI integration. This conceptual positioning preserves human responsibility and judgment as the primary locus of meaning in human–AI interaction. While serving as the theoretical base for future applications, the paper defines the ethical landscape for responsible academic research.

KEYWORDS: AI Ethics, Epistemic Archaeology, Sociotechnical Pragmatism, Human–AI Interaction, Knowledge Production, Epistemic Literacy

Introduction

The increasing presence of artificial intelligence within academic research has progressively transformed practices of knowledge production, especially in the humanities and social sciences. Generative systems are now routinely employed for exploratory literature searches, conceptual mapping, translation and preliminary synthesis. This diffusion has prompted renewed attention to questions of academic ethics, including authorship attribution, epistemic reliability, transparency of sources and the role of human responsibility in mediated forms of research.

A growing body of literature suggests that both human cognition and artificial intelligence are shaped by preexisting social and cultural conditions. According to Ho and Vuong (2025, p. 1), neither humans nor machines operate independently from historical, political and cultural structures that predispose ways of reasoning and interpretation. From this standpoint, AI systems inherit epistemic orientations embedded in the corpora on which they are trained, rather than producing knowledge in a neutral or context-free manner. This observation is particularly relevant for academic environments, where interpretive sensitivity and theoretical positioning are central to academic rigor.

Recent debates have also emphasized that artificial intelligence should not be treated as a monolithic entity. Narayanan and Kapoor (2024) underline that the label “AI” includes a heterogeneous range of techniques and architectures, each characterized by distinct affordances and limitations. Among these, large language models have gained prominence due to their capacity to generate coherent and stylistically convincing text. However, their operational logic remains fundamentally statistical. As Bender et al. (2021) observe, such systems rely on the identification of distributional patterns in language rather than on any form of grounded understanding of the world.

This distinction carries important epistemic consequences. While human researchers can suspend judgment, reflect on uncertainty and revise their interpretations, language models are structurally designed to produce an output for every prompt. Quattrociocchi, Capraro and Perc (2025, p. 4) note that LLMs reconstruct meaning indirectly through textual correlations, lacking the experiential and contextual grounding that characterizes human cognition. As a result, fluency and plausibility may hide the absence of epistemic validation, particularly when outputs are read uncritically.

Concerns related to bias and cultural representation further complicate the use of generative AI in academic contexts. Empirical studies indicate that many models disproportionately reflect values and moral intuitions typical of Western, educated and industrialized societies. According to Atari et al. (2023, p. 2), this imbalance is largely attributable to the dominance of English-language and Western-centric datasets. Similarly, Blasi et al. (2022) show that linguistic diversity remains unevenly represented in large-scale language technologies, with potential consequences for the visibility of non-dominant epistemic traditions.

At the same time, the complete exclusion of AI tools from research practices appears neither feasible nor desirable. Floridi et al. (2018) argue that computational systems increasingly participate in scientific discovery processes, shaping how problems are framed and explored. In this sense, the ethical challenge is not limited to the risks of misuse but concerns the broader conditions under which AI-mediated knowledge is produced and interpreted. As Ho and Vuong (2025, p. 2) point out, values emerge through interactions between humans and technologies, becoming embedded in social practices regardless of explicit normative reflection.

Within this context, sociotechnical pragmatism offers a useful interpretive lens. Rather than evaluating AI systems in isolation, this approach emphasizes their role within specific institutional, cultural and epistemic arrangements. Watson et al. (2025, p. 3170) suggest that ethical assessment should consider not only what AI systems can do, but also how their use compares with available alternatives and how responsibility is distributed among human actors. Pragmatism further highlights the importance of responsibility, understood as humans taking ownership of their claims and actions within complex sociotechnical environments (Brandom, 1979; Rorty, 2021).

Building on these theoretical premises, this article introduces the notion of epistemic archaeology as a conceptual framework for examining AI-assisted research practices. Rather than treating AI outputs as authoritative contributions, epistemic archaeology approaches them as stratified artifacts shaped by statistical regularities, cultural traditions and implicit worldviews. Differences between models are thus interpreted not merely as technical variance, but as signals of underlying epistemic orientations. In this sense, AI systems become objects of critical inquiry rather than sources of epistemic delegation.

This article is positioned as a foundational conceptual study; therefore, while it provides the theoretical infrastructure for integrative methodologies (such as AI-enhanced systematic reviews) specific procedural workflows and empirical operationalizations lie outside its current scope. By focusing on the ethical and epistemological coordinates of AI integration, the analysis seeks to establish a 'critical infrastructure' for governing mediated knowledge. This theoretical grounding is a necessary prerequisite for subsequent applied developments, ensuring that practical tools remain anchored in interpretive responsibility and epistemic pluralism. The goal is to provide a robust conceptual orientation for engaging with GenAI as a situated participant in scholarly production

AI, Science and Epistemic Transformation

This paper argues that within contemporary scientific inquiry, the integration of artificial intelligence entails transformations that concern epistemic structures rather than merely operational processes. While early discussions emphasized gains in speed, scale and efficiency, current applications indicate a deeper transformation in how knowledge is explored, organized and preliminarily articulated. Large language models are now routinely involved in activities such as literature triage, exploratory synthesis, hypothesis generation and the coordination of complex discovery pipelines. As documented by Song et al. (2025, p. 2), LLMs already demonstrate "promise in a great variety of scientific discovery projects", particularly in contexts where exploration and serendipity play a central role. Similarly, recent work highlights their increasing use across multiple stages of scientific workflows, from the early mapping of research landscapes to downstream computational tasks (Vaswani et al., 2017; Jablonka et al., 2024).

This expansion of AI-mediated practices does not simply accelerate established forms of inquiry. Rather, it alters the epistemic ecology in which scientific reasoning unfolds. By intervening at the level of exploration, AI systems shape which literatures are foregrounded, which connections are rendered salient and which hypotheses appear plausible at an early stage. In doing so, they influence the pre-evaluative space of research, where questions are framed and directions are selected before formal validation occurs. The transformation at stake therefore concerns the organization of epistemic attention, not only the automation of tasks.

At the core of this transformation is a fundamental distinction between human inquiry and machine generation. Large language models do not pursue

truth, explanation or understanding in any substantive sense. Their operation consists in optimizing statistical regularities within vast textual datasets. As has been widely noted, LLMs are trained on human generated language and learn to estimate the probability of token sequences rather than to represent states of the world (Bender et al., 2021). Quattrociocchi, Capraro and Perc (2025, p. 3) describe this process with precision, emphasizing that each output corresponds to “the realization of a stochastic trajectory generated by local sampling” within a high dimensional state space.

From this perspective, textual generation must be understood as an ergodic and probabilistic process. It explores a space of linguistic possibilities under statistical constraints, rather than converging toward epistemic correctness. Text generation “is therefore an ergodic process under statistical constraints, not a procedure of epistemic convergence” (Quattrociocchi, Capraro and Perc, 2025, p. 3). What appears as a conclusion is not the endpoint of evaluation, but the terminal state of a probabilistic sequence. This property has significant consequences when AI systems are incorporated into scientific contexts, where conclusions conventionally signal the outcome of cumulative assessment.

Human scientific judgment, by contrast, is embedded in practices of evaluation that presuppose world directedness, uncertainty monitoring and the possibility of withholding commitment. Humans ground interpretation in perceptual experience, social context and theoretical frameworks that are continuously revised through interaction with evidence. They rely on a layered, meaning rich model of the environment, whereas LLMs reconstruct meaning indirectly from text alone. The difference is not one of degree, but of kind. One system parses a world; the other segments a string.

The ethical meaning of this distinction becomes apparent when AI generated outputs are introduced into epistemic practices that rely on markers of authority and credibility. Because LLM outputs are typically fluent, coherent and stylistically aligned with academic discourse, they may be perceived as epistemically vetted even when no such validation has occurred. The primary risk, therefore, is not simple error. Errors are an ordinary feature of scientific inquiry and are often productive. The more subtle risk lies in the attribution of epistemic authority to outputs that have not undergone evaluation. Quattrociocchi, Capraro and Perc (2025, p. 9) observe that these systems are capable of producing “rhetorically convincing outputs without performing epistemic evaluation”.

Empirical evidence from AI assisted discovery reinforces this concern. Song et al. (2025, p. 13) show that top performing models from different providers often converge on identical incorrect answers when confronted with difficult scientific questions. This convergence suggests that agreement across models may reflect shared training distributions rather than independent epistemic justification. In such cases, apparent consensus can obscure systematic limitations rooted in data homogeneity and statistical optimization. The problem is not that models are wrong, but that their errors may appear collectively reinforced.

These dynamics must be situated within broader sociotechnical conditions. AI systems are trained on data drawn from preexisting social worlds, and their outputs inevitably reflect the structures, priorities and omissions embedded in those worlds. As Ho and Vuong (2025, p. 1) argue, algorithms that interact with humans are necessarily shaped by the social and cultural environments from which their data originate. When deployed in scientific research, LLMs therefore participate in the reproduction and reconfiguration of epistemic norms, often without explicit reflection.

The transformation of science under AI mediation thus concerns the redistribution of epistemic functions. AI systems expand the exploratory horizon by rapidly traversing large textual spaces and proposing plausible configurations of knowledge. Humans retain responsibility for evaluation, justification and commitment. However, this division of labor is fragile. When exploratory outputs are treated as epistemic contributions rather than as heuristic artifacts, the boundary between exploration and validation becomes blurred. The observation has already been made that values emerging from human-computer interactions can be incorporated into social practices regardless of whether they are consciously examined (Ho & Vuong, 2025, p. 2). In scientific contexts, this incorporation can affect standards of rigor, authorship and accountability.

Understanding AI as an epistemic transformer rather than a neutral accelerator clarifies the ethical stakes of its use in research. The challenge is not to eliminate AI from scientific practice, but neither to treat its outputs as authoritative. It is to recognize that generative systems reshape how knowledge is explored while remaining incapable of epistemic judgment. This recognition sets the stage for a sociotechnical evaluation of AI assisted research, which must attend to practices, contexts and responsibilities rather than isolated outputs. The following section develops sociotechnical pragmatism as a framework for articulating such an evaluation.

Sociotechnical Pragmatism as an Ethical Framework

The epistemic transformation described in the previous section calls for an ethical framework capable of accounting for the situated and relational nature of AI assisted research practices. Evaluative approaches that focus on individual system outputs or on isolated technical properties remain insufficient when generative models are embedded in academic workflows, institutional routines and interpretive communities. Sociotechnical pragmatism offers a suitable response to this challenge by emphasizing the inseparability of technical artifacts from the social, cultural and organizational contexts in which they operate.

Within this perspective, artificial intelligence is not treated as an autonomous epistemic agent but as a component of broader sociotechnical systems. The reference literature emphasizes that algorithms inevitably function within configurations that include people, norms, infrastructures and organizations, and individual components of such systems “cannot be analyzed in isolation” (Lazar and Nelson, 2023; Leveson, 2017 p. 560). Ethical concerns

therefore arise not solely from the internal functioning of models, but from the ways in which they are integrated into practices of inquiry, evaluation and decision making. Lauer (2021) similarly notes that many AI related harms stem from the dynamic interactions between system components and their uses, rather than from discrete technical failures.

This systemic orientation aligns with the definition of sociotechnical pragmatism proposed by Watson et al. (2025, p. 3163) as an approach that foregrounds context and human agency in both the design and evaluation of emerging technologies. Rather than relying on fixed normative criteria, this framework acknowledges that value judgments are irreducibly context dependent and must be negotiated in relation to specific purposes and constraints. Watson et al. (2025, p. 3164) explicitly argue for an ethical stance that is principled yet flexible, relational rather than relativist, and attentive to feedback, learning and redress within concrete institutional settings.

From this standpoint, ethical evaluation cannot be reduced to the assessment of single outputs. In academic research, generative AI systems increasingly participate in activities such as literature exploration, conceptual organization and hypothesis articulation. Their influence emerges through repeated interaction rather than through isolated results. This observation resonates with pragmatist accounts of responsibility, according to which freedom consists in the capacity of agents to take responsibility for their claims and actions within shared normative practices (Brandom, 1979; Rorty, 2021). Applied to AI assisted research, this implies that responsibility remains a human achievement sustained through interpretive judgment, justification and revision, rather than something that can be delegated to computational systems.

A further implication of sociotechnical pragmatism concerns the question of non-use. Floridi et al. (2018) caution that abstaining from machine learning systems altogether may incur significant social opportunity costs. In research contexts, refusing AI mediated tools may constrain exploratory capacity and limit engagement with emerging forms of scientific discovery. At the same time, uncritical adoption carries its own risks. Watson et al. (2025, p. 3164) warn that machine learning systems can “codify and amplify injustices already present in society”, a concern echoed by analyses of how automated systems reproduce existing social inequalities (McGregor, 2021). Ethical evaluation must therefore operate in the space between these alternatives, comparing concrete practices rather than adopting absolute positions.

This comparative orientation is central to sociotechnical pragmatism. Normative assessment cannot be conducted in isolation but only in relation to available alternatives and institutional arrangements. Transparency about normative tradeoffs becomes part of ethical practice, informing both internal deliberation among researchers and broader public discussions about the values embedded in scientific infrastructures (Watson et al., 2025, p. 3172). Such transparency is particularly relevant in AI assisted research, where the apparent

neutrality of computational tools may obscure underlying assumptions and exclusions.

Pluralism plays a crucial role within this framework. Page (2018) argues that engaging with multiple models is essential for making sense of complex social phenomena, since no single model can capture all relevant dimensions. Sociotechnical pragmatism adopts this insight by treating plurality not as a source of confusion but as a resource for inquiry. Watson et al. (2025, p. 3169) describe pragmatism as rooted in multi model thinking, providing a robust stance for assembling heterogeneous forms of evidence and interrogating them through complementary lenses. In this view, divergence across models is epistemically informative rather than merely problematic.

This pluralistic orientation has direct ethical implications for AI assisted research practices. Employing multiple generative models does not primarily serve the goal of identifying the most accurate or authoritative output. Instead, it enables researchers to examine how different systems, shaped by distinct training data and design choices, foreground or marginalize particular concepts, assumptions and interpretive frames. Such an approach treats AI systems as exploratory instruments whose differences invite critical comparison. Ethical responsibility is thus exercised not through reliance on a single system, but through the deliberate governance of interaction across heterogeneous tools.

Sociotechnical pragmatism conceptualizes this process as exploratory problem solving, understood as a collective practice in which communities examine empirical situations in order to improve conditions they have identified as requiring change (Prasad, 2021). Within academic research, this implies that AI assisted methods should remain open to revision and contextual adjustment, supported by institutional norms that sustain accountability and reflexivity. The ethical focus shifts from evaluating individual outputs to shaping practices that preserve human judgment as the locus of epistemic responsibility.

In this sense, sociotechnical pragmatism provides the ethical foundation for integrative approaches to AI assisted research. By foregrounding context, pluralism and responsibility, it reframes the use of generative models as a means of interrogating epistemic complexity rather than outsourcing judgment. This orientation prepares the ground for methodologies that leverage multiple AI systems as part of a controlled and reflective research process, while maintaining the autonomy and accountability of the human researcher.

LLMs as Cultural and Psychological Objects

Large language models are often described in functional terms, as tools for prediction, generation, or automation. Such descriptions, while technically accurate, risk obscuring a more consequential dimension of these systems: their status as cultural and psychological artifacts. LLMs are trained on vast corpora of human generated text, and through this process they internalize patterns that reflect not only linguistic regularities but also culturally situated forms of cognition, valuation, and social interaction. Building on these empirical insights,

I propose that understanding LLMs as cultural and psychological objects is therefore essential for any ethical assessment of their role in academic research.

Atari et al. (2023, p. 2) observe that much of the existing literature on large language models overlooks a fundamental fact: humans are a cultural species characterized by substantial psychological diversity across populations, a diversity that is not adequately represented in the textual data used to train contemporary models. This omission is not a marginal limitation but a structural feature of current LLM development. The psychological profiles encoded in these systems are shaped by the distributions of voices, values, and perspectives present in their training data, which are themselves unevenly drawn from global populations. As a result, LLMs cannot be assumed to reflect a neutral or universal model of human cognition.

This point gains further clarity when LLMs are approached, following Horton (2023), as "implicit computational models of humans". Under this interpretation, language models do not merely manipulate symbols but enact probabilistic approximations of how humans reason, evaluate, and communicate. However, the humans implicitly modeled are not humanity in its plural forms. Mainstream research and development practices disproportionately reflect populations from Western, Educated, Industrialized, Rich, and Democratic contexts. Atari et al. (2023, p. 4) note that psychological dimensions such as moral reasoning, cooperation, trust, and self-concept vary significantly across cultures, yet these variations remain largely absent from the data regimes that shape LLM behavior.

The consequence is what the literature has described as a WEIRD bias embedded within generative models. According to Atari et al. (2023, p. 16), LLMs inherit a WEIRD psychology across both attitudinal domains, including values and belief systems, and cognitive domains, such as thinking style and self-perception. This inheritance is not accidental. It follows directly from the fact that the textual material used for training is predominantly produced by individuals from WEIRD populations. Bender et al.'s (2021) characterization of LLMs as "stochastic parrots" captures this dynamic with particular force, emphasizing that these systems recombine existing linguistic patterns without grounding them in lived experience or world directed understanding. When the source material itself is culturally skewed, the resulting outputs reproduce and amplify that skew.

Language plays a central role in this process. Blasi et al. (2022) demonstrate that English is overwhelmingly dominant in language technologies, far exceeding its proportional representation among the world's languages. This dominance has epistemic implications that extend beyond communication efficiency. Language encodes culturally specific categories, metaphors, and normative assumptions. When English functions as the primary medium through which models learn to represent the world, alternative epistemic traditions and conceptual frameworks are systematically underrepresented. The resulting models exhibit a form of linguistic centralization that parallels, and reinforces, broader asymmetries in knowledge production.

Opacity further complicates this picture. LLMs are trained on massive datasets through architectures whose internal representations are not directly interpretable. As Atari et al. (2023, p. 5) point out, the psychology that these models acquire from their training data and apply to downstream tasks remains largely unknown due to this opacity. While the models' outputs may appear coherent and contextually appropriate, the processes through which cultural assumptions are encoded and operationalized are not accessible to scrutiny in any straightforward sense. This limits the capacity of researchers to identify, contest, or correct embedded biases at the level of internal model dynamics.

The ethical significance of these observations becomes clearer when situated within a broader understanding of human cognition. Humans are not uniform reasoning machines but culturally evolved agents shaped by social learning and cognitive plasticity (Boyd et al., 2011). Even processes sometimes labeled as basic or low-level exhibit substantial variation across populations, as Atari et al. (2023, p. 4) emphasize in their discussion of cross-cultural differences in cognition. Treating LLM outputs as representative of "human like" reasoning therefore risks collapsing a complex landscape of psychological diversity into a narrow and historically contingent profile.

From an epistemic perspective, this collapse has direct consequences for academic research. When LLMs are employed to assist in literature exploration, conceptual framing, or hypothesis articulation, they do not merely accelerate existing practices. They introduce patterned tendencies shaped by the cultures embedded in their training data. These tendencies influence which arguments appear salient, which theoretical traditions are foregrounded, and which forms of reasoning are rendered intuitive or marginal. The authority of fluent language generation can mask these influences, making culturally specific assumptions appear as general insights.

Bender et al. (2021) caution that large language models operate on statistical regularities extracted from human produced text rather than on representations of the world itself. This distinction matters because it clarifies that apparent coherence does not entail epistemic grounding. Quattrociocchi, Capraro, and Perc (2025, p. 3) further argues that each output generated by an LLM constitutes the realization of a stochastic trajectory within a constrained state space, optimized for distributional fit rather than correctness with respect to the world. When such trajectories are shaped by culturally skewed data, the resulting outputs reflect a specific epistemic lineage rather than an impartial synthesis.

The ethical implication is not that LLMs are inherently flawed or unusable, but that there is no single entity that can be meaningfully described as "the AI." Instead, there exists a plurality of models, each trained on particular corpora that encode specific cultural, linguistic, and psychological configurations. These models differ not only in architecture or performance metrics, but in the implicit conceptions of humanity they enact. Ignoring this plurality risks attributing unwarranted universality and authority to systems whose perspectives are, in fact, partial.

Recognizing LLMs as cultural and psychological objects reframes the ethical task. Rather than asking whether a given model is biased in the abstract, the relevant question becomes which human experiences and epistemic traditions are being represented, which are absent, and how these distributions shape research practices. Atari et al. (2023, p. 16) explicitly call for continuous reflection on “which humans” serve as the sources of training data in generative models, emphasizing the need for both scientific and industrial efforts to diversify data and annotation practices. This call underscores that ethical evaluation cannot be separated from questions of representation and inclusion.

This is essential to establish a critical premise for the analysis that will be carried out in the next part. If LLMs embody distinct cultural psychologies, then divergences between models are not mere technical noise. They are traces of different epistemic histories encoded in data. Attending to these divergences requires interpretive strategies capable of reading model variation as meaningful rather than accidental. Such a perspective opens the way for an archaeological approach to AI assisted knowledge production, one that treats generative outputs as layered artifacts shaped by sedimented cultural influences. It is precisely this interpretive move that the next part develops through the concept of epistemic archaeology.

From Epistemic Transformation to Epistemic Archaeology

The increasing integration of large language models into academic research practices requires a conceptual framework capable of accounting not only for their functional performance but also for the epistemic conditions under which their outputs are produced. Building on the sociotechnical and cultural framework previously discussed, this paper develops 'epistemic archaeology' as an original interpretive extension. This practice is oriented toward the systematic reading of AI-generated outputs as stratified epistemic artifacts, shifting the focus from output validation to the analysis of the conditions that make specific responses probable. Rather than treating variation across models as an obstacle to reliability or as a source of noise to be minimized, epistemic archaeology advances the claim that such variation constitutes a meaningful signal. Differences between outputs are neither accidental nor merely technical. They are traces of distinct training corpora, statistical priors, institutional choices, and cultural traditions embedded in the development of each model.

From this perspective, every response generated by an LLM can be understood as the contingent realization of a specific epistemic history. Prior to any individual system, there exist social, political, cultural, and historical structures that predispose both humans and machines toward particular modes of reasoning and expression. As Ho and Vuong (2025, p. 1) observe, algorithms that interact with humans are trained on data originating from preexisting social worlds. Epistemic archaeology takes this observation seriously but shifts its analytical focus from abstract ethical assessment to close interpretive scrutiny of

outputs themselves. What matters is not only that models inherit biases, but how those biases become legible in language, framing, and argumentative structure.

Standard evaluation paradigms in AI research often assume that epistemic reliability increases as outputs converge. Benchmarks are designed to reward alignment with a single correct answer, and disagreement between systems is frequently treated as a deficiency. However, recent empirical findings complicate this assumption. Song et al. (2025, p. 13) report that top-performing models from different providers often exhibit high correlations in their errors, converging on identical incorrect answers in the most challenging tasks. This pattern suggests that convergence may reflect shared epistemic limitations rather than independent verification. In such cases, similarity is not evidence of truth but an indicator of common training distributions and overlapping epistemic constraints.

Epistemic archaeology reinterprets this phenomenon. When multiple models fail in the same way, the failure itself becomes informative. It points to dominant assumptions embedded in large-scale textual corpora, to disciplinary orthodoxies reproduced through data selection, and to institutional priorities shaping what counts as relevant or authoritative knowledge. Conversely, when models diverge in their responses to identical prompts, these divergences illuminate alternative epistemic orientations. Differences in emphasis, normative framing, or conceptual vocabulary reveal how distinct corpora and optimization strategies encode particular visions of the world.

A crucial extension of epistemic archaeology concerns the systematic analysis of what appears only in some models and not in others. Divergence is not limited to stylistic variation or differences in argumentative emphasis. In many comparative settings, identical prompts elicit the emergence of themes, analogies, normative framings, or problem definitions that are present in one model and entirely absent in another. These asymmetries are particularly revealing because they cannot be explained by stochastic fluctuation alone. Rather, they function as epistemic markers pointing toward the selective presence or absence of entire discursive traditions within training corpora.

When a model consistently introduces specific conceptual references, moral concerns, or explanatory frameworks that do not surface elsewhere, such outputs should be read as traces of differential exposure to textual domains. A recurring emphasis on legal compliance, for instance, may reflect training data heavily saturated with regulatory discourse. Similarly, the spontaneous invocation of geopolitical risk, environmental externalities, or social justice narratives may signal the prominence of policy-oriented, activist, or institutional sources within the corpus. Conversely, the systematic absence of certain perspectives can be equally informative, revealing zones of epistemic silence shaped by data exclusion, linguistic dominance, or commercial priorities.

From an archaeological standpoint, these unique or weakly correlated themes function analogously to rare artifacts in a stratigraphic excavation. Their analytical value lies not in frequency but in specificity. While high error correlations across frontier models suggest shared epistemic constraints imposed

by overlapping data distributions (Song et al., 2025, p. 13), thematic divergence exposes the residual heterogeneity that remains despite large-scale convergence. Epistemic archaeology attends precisely to this residue, treating it as a site of epistemic inquiry rather than as noise to be suppressed.

This focus on divergence also counters a common tendency in AI evaluation to privilege consensus over interpretive richness. When models agree, agreement is often taken as validation. Yet, as discussed earlier, convergence may reflect shared blind spots rather than independent corroboration. Divergent themes, by contrast, can surface alternative framings of the same problem, revealing how different epistemic cultures prioritize values, risks, or causal explanations. In this sense, divergence does not undermine reliability but expands the epistemic horizon within which judgment can be exercised.

Engaging productively with such variation requires abandoning the idea that one model must be selected as epistemically superior. Instead, epistemic archaeology aligns with the insight that engaging with multiple models is essential for understanding complex phenomena. According to Page (2018), multi-model thinking enables analysts to capture dimensions of reality that remain inaccessible when relying on a single perspective. Applied to generative AI, this approach reframes model plurality as an epistemic resource rather than a technical inconvenience. Each model becomes a partial testimony to the epistemic culture that produced it.

This interpretive stance also responds to a fundamental limitation of LLMs themselves. Large language models operate on statistical regularities extracted from human-produced text rather than on representations of the world. Text generation is therefore an ergodic process under statistical constraints, not a procedure of epistemic convergence. Each output represents the terminal state of a stochastic trajectory, optimized for distributional fit rather than for correctness with respect to external reality. As Quattrociocchi, Capraro and Perc (2025, p. 3) state, “a conclusion is not the terminus of evaluation, but the terminal state of a stochastic trajectory”.

This operational logic has profound epistemic consequences. Humans ground judgments in perceptual reality, social context, and lived experience, whereas LLMs must reconstruct meaning indirectly from text alone. Moreover, humans possess a metacognitive layer that allows them to monitor uncertainty, detect potential errors, and, when necessary, withhold judgment. LLMs possess no such capacity. Humans can refrain from judging, whereas generative systems must always produce a prediction (Quattrociocchi, Capraro and Perc, 2025, p. 7). The absence of metacognition means that LLMs cannot evaluate the epistemic adequacy of their own outputs. Responsibility for interpretation and validation is therefore displaced onto human users.

Epistemic archaeology emerges precisely at this point of displacement. It provides a framework for assuming interpretive responsibility without attributing epistemic authority to the system itself. Rather than asking whether an AI-generated response is correct in isolation, the archaeological approach asks what

conditions made this response likely. Each output is treated as a layered artifact, shaped by probabilistic inference, cultural regularities, and institutional decisions embedded in training pipelines. Differences across models are read comparatively, allowing researchers to identify recurring patterns, shared blind spots, and distinctive epistemic signatures.

This approach is particularly relevant given evidence that current frontier models may be approaching performance levels imposed not by architecture but by data homogeneity. Song et al. (2025, p. 13) suggest that the convergence of errors across models points toward limitations rooted in similar pre-training distributions. From an archaeological standpoint, such plateaus are not merely technical challenges but epistemic indicators. They reveal the boundaries of dominant knowledge infrastructures and highlight the need to interrogate the sources from which models learn.

Crucially, epistemic archaeology does not deny the practical value of LLMs in scientific discovery. Large language models have already demonstrated their capacity to accelerate literature triage, hypothesis generation, and exploratory reasoning (Song et al., 2025, p. 2). However, acceleration without interpretation risks reinforcing existing epistemic asymmetries. When models trained on similar corpora produce similar outputs, apparent consensus may conceal structural bias rather than reflect robust knowledge. Archaeological analysis resists this illusion by foregrounding difference as an object of inquiry.

The implications of this framework extend beyond methodology to questions of power and governance. Proprietary models are trained on datasets curated within specific geopolitical, economic, and regulatory contexts. These contexts shape not only data availability but also normative priorities regarding safety, inclusiveness, and acceptable risk. As a result, there is no singular AI perspective but a plurality of model-specific articulations, each reflecting a particular configuration of social and institutional forces. Treating AI-generated content as homogeneous obscures this plurality and encourages unwarranted trust in apparent neutrality.

Epistemic archaeology insists instead on asking which epistemic cultures speak through each model. It treats outputs as traces of training materials and design decisions that are often opaque but not inaccessible to interpretation. By systematically comparing responses across models, researchers can infer the contours of these underlying epistemic formations, much as archaeologists reconstruct past societies from material remains. In this sense, AI outputs function as contemporary artifacts, bearing the imprint of the cultural worlds that produced them.

Ultimately, epistemic archaeology reframes the role of AI in knowledge production. Generative systems are neither authoritative sources nor passive tools. They are epistemic objects whose outputs require interpretation, contextualization, and comparative analysis. Preserving judgment as an accountable human practice in an environment increasingly organized around generative AI requires more than improved models. It requires interpretive

frameworks capable of making sense of difference without collapsing it into error. Epistemic archaeology offers such a framework, positioning plurality not as a problem to be solved but as a condition to be understood.

Humanities and Social Sciences as Critical Mediators in AI-Mediated Knowledge

The humanities and social sciences frequently emerge as the domains most exposed to automation in narratives surrounding the expanding role of LLMs in research, policy, and cultural production. Within this framing, interpretive disciplines are portrayed as increasingly redundant in the face of systems capable of generating text, synthesizing literature, and simulating reasoning at unprecedented scale. This paragraph advances a counter-claim. Far from being marginalised by artificial intelligence, the humanities and social sciences constitute the critical infrastructures that make AI-mediated knowledge epistemically governable. Their role is not to compete with computational systems on speed or volume, but to mediate interpretation, provide context, assign responsibility, and cultivate epistemic literacy in hybrid human–AI environments.

This repositioning requires abandoning the assumption that AI systems function as epistemically self-sufficient agents. As discussed in previous chapters, large language models operate through statistical inference over human-generated text, optimizing token prediction without access to the world they describe. Their internal processes, grounded in layered neural computation, propagate representations through attention mechanisms and nonlinear transformations in order to minimize predictive error (Goodfellow, Bengio, & Courville, 2016; Vaswani et al., 2017). This optimization-driven architecture enables impressive generative fluency, yet it remains indifferent to truth, value, or consequence. Meaning, relevance, and normative weight are not properties computed by the model but effects attributed by human interpreters.

From this perspective, the humanities and social sciences do not intervene as external ethical correctives applied after technical development. They operate as mediating practices embedded within the epistemic lifecycle of AI use. Interpretation is the first of these mediations. AI-generated outputs are not transparent representations of knowledge but probabilistic realizations shaped by training data, architectural constraints, and optimization objectives. As Dasgupta et al. (2022) argue, the linguistic regularities learned by such models reflect not only human language use but also the structural biases of model design, which in turn influence how reasoning tasks are approached. Interpreting outputs therefore requires sensitivity to framing, omission, metaphor, and rhetorical structure; competencies long cultivated within humanistic inquiry.

Contextualization constitutes a second, closely related function. Outputs produced by LLMs are context-poor in an epistemic sense, even when they appear contextually fluent. They reconstruct meaning indirectly from textual associations rather than grounding it in lived experience, institutional practice, or social consequence. Human judgment, by contrast, is situated. It integrates perceptual

reality, social norms, historical memory, and pragmatic goals. Motivation and goal orientation structure human decision-making by defining what counts as success or failure within a given context (Locke & Latham, 2002). LLMs possess no equivalent teleological structure. Their outputs are generated without intrinsic aims, emotional salience, or awareness of stakes. Humanities and social sciences supply the contextual scaffolding that allows AI-generated material to be situated within broader interpretive and institutional frameworks.

Responsibility represents a third axis of mediation. AI systems cannot be held accountable for the judgments they appear to make, because they do not possess metacognitive capacities for uncertainty monitoring, error detection, or value-based deliberation. They must always produce an output, regardless of ambiguity or risk. Humans, by contrast, retain the capacity to suspend judgment, demand justification, and revise beliefs in light of new evidence. As Quattrociocchi, Capraro and Perc (2025, p. 2) emphasize, preserving judgment as an accountable human practice in hybrid systems requires more than improved technical alignment; it requires institutional and cultural mechanisms through which reasons can be demanded and errors owned. The humanities and social sciences contribute precisely to these mechanisms by sustaining traditions of critique, reflexivity, and normative debate.

This mediating role becomes particularly salient when considering the power effects of AI deployment. AI systems are not neutral instruments but components of sociotechnical arrangements that can reconfigure power relations, often in ways that advantage some groups over others (Lazar, 2022). Decisions about data curation, model deployment, and application domains embed political and economic priorities that shape whose knowledge is amplified and whose is marginalised. Pragmatist approaches to governance have long emphasized that technological solutions must be complemented by legal, cultural, and institutional interventions if they are to serve democratic ends (Maas, 2022). Humanities and social sciences provide the analytical tools necessary to surface these power dynamics and to assess their social implications.

Epistemic literacy constitutes the fourth and integrative function of mediation. In environments increasingly organized around generative AI, epistemic competence can no longer be reduced to the evaluation of arguments or sources in isolation. It must encompass the ability to decide when to rely on generative tools, how to interpret their outputs, and how to govern the division of epistemic labor between human and machine. Quattrociocchi, Capraro and Perc (2025, p. 10) frame this as an extension rather than a replacement of critical thinking: epistemic literacy complements traditional critique by addressing the governance of judgment in hybrid systems. Humanities and social sciences are uniquely positioned to cultivate this literacy, drawing on methodologies that combine qualitative interpretation, historical analysis, and normative reasoning.

Within this framework, the relationship between AI and human judgment is not conflictual but complementary. Computational systems excel at amplifying exploration. They can scan vast literatures, generate candidate hypotheses, and

surface patterns that would remain inaccessible to individual researchers. Human judgment, however, governs meaning. It determines which questions are worth asking, which outputs are relevant, and which interpretations carry normative weight.

This stance resists both technological solutionism and humanist exceptionalism. Abstaining from the use of machine learning systems altogether may entail significant social opportunity costs, particularly in domains where scale and complexity exceed human capacities (Floridi et al., 2018). At the same time, uncritical adoption risks codifying and amplifying existing inequalities, as systems trained on historically skewed data reproduce dominant perspectives while silencing others. The mediating role of the humanities and social sciences lies in navigating this tension; neither rejecting AI as an external threat nor embracing it as an epistemic authority but integrating it within practices of responsible inquiry.

Seen in this light, the humanities are not displaced by AI-generated text; they are activated by it. Every output becomes an object of interpretation, a site where cultural assumptions, institutional priorities, and statistical regularities intersect. Social sciences, similarly, are not rendered obsolete by automated analysis; they are required to situate such analysis within social structures, power relations, and collective goals. Together, these disciplines function as critical infrastructures that enable societies to benefit from AI's exploratory capacities without relinquishing control over meaning, value, and responsibility.

In conclusion, the future of AI-assisted knowledge production does not hinge on replacing human judgment but on redistributing epistemic labor in a way that preserves accountability. Humanities and social sciences play a central role in this redistribution. They mediate between computation and consequence, between probabilistic generation and normative evaluation. By doing so, they transform AI from a disruptive force into a governable component of collective inquiry, ensuring that acceleration does not come at the expense of understanding, and that exploration remains anchored in human responsibility.

Conclusion

This paper articulates a conceptual positioning of artificial intelligence within contemporary debates on academic ethics and the future of the humanities and social sciences. Rather than advancing a technical proposal or a procedural framework, the discussion establishes the epistemic and ethical conditions under which AI-assisted research can be meaningfully integrated into scholarly inquiry. By situating AI within broader sociotechnical, cultural, and epistemic structures, the article emphasizes complexity, plurality, and responsibility as central analytical requirements.

The analysis shows that the growing involvement of large language models in scientific practices transforms how knowledge is explored, framed, and rendered visible. These systems intervene in the organization of epistemic attention, shaping the selection of sources, the articulation of research problems,

and the generation of possible explanatory trajectories. At the same time, their operational logic remains grounded in probabilistic optimization over textual regularities. Text generation unfolds as an ergodic process constrained by statistical distributions, producing outputs that are coherent and rhetorically persuasive without being subjected to epistemic evaluation. The ethical issue therefore does not primarily concern isolated inaccuracies, but the risk that plausibility is misrecognized as epistemic authority.

Understanding AI systems as cultural and psychological artifacts further clarifies this risk. Large language models inherit linguistic, cultural, and cognitive patterns embedded in their training data, which are unevenly distributed across populations, languages, and epistemic traditions. Their outputs thus reflect situated worldviews rather than neutral representations of human knowledge. Acknowledging this partiality becomes essential for resisting reductive interpretations of AI as an objective or universal epistemic agent.

Within this context, the notion of epistemic archaeology provides an interpretive orientation that treats variation across models as analytically meaningful. Divergences, convergences, and absences emerging from comparisons between multiple models responding to identical prompts can be read as traces of underlying epistemic conditions, including dataset composition and representational biases. Such an approach does not displace established scholarly methodologies but operates alongside them, extending the analytical space in which AI-assisted outputs are examined.

The role of the humanities and social sciences emerges, in this perspective, not as peripheral or defensive, but as infrastructural. Through practices of interpretation, contextualization, responsibility, and epistemic literacy, these disciplines sustain the conditions under which judgment remains accountable to evidence, values, and social consequences. Integrative methodological orientations grounded in systematic analysis and enhanced by AI tools exemplify this stance. In these configurations, generative systems expand the exploratory capacity of research, while the attribution of meaning, relevance, and validity remains a distinctly human responsibility.

In an epistemic environment increasingly shaped by generative technologies, academic ethics cannot be reduced to technical safeguards or model refinement alone. It requires preserving the social and institutional practices through which reasons can be demanded, interpretations contested, and errors acknowledged. From this standpoint, artificial intelligence does not signal the marginalization of the humanities and social sciences, but rather reinforces their function as critical mediators in the governance of contemporary knowledge production.

Acknowledgements

The author gratefully acknowledges the financial support of the European Union – NextGenerationEU and the Italian Ministry of University and Research (MUR) through the PNRR (Mission 4, Component 2, Investment 3.3), within the framework of the DM 630 doctoral program at the University of Pavia. Project CUP: F13C24000390005.

REFERENCES:

Atari, M., Xue, M. J., Park, P. S., Blasi, D. E., & Henrich, J. (2023). Which Humans? *PsyArXiv*. https://doi.org/10.31234/osf.io/5b26t

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*, 26(12), 1153–1170. https://doi.org/10.1016/j.tics.2022.09.015

Boyd, R., Richerson, P. J., & Henrich, J. (2011). The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences*, *108 (supplement_2)*, 10918–10925. https://doi.org/10.1073/pnas.1100290108

Brandom, R. (1979). Freedom and Constraint by Norms. *American Philosophical Quarterly*, 16(3), 187–196.

Dasgupta, I., Lampinen, A. K., Chan, S. C. Y., Sheahan, H. R., Creswell, A., Kumaran, D., McClelland, J. L., & Hill, F. (2022). Language models show human-like content effects on reasoning tasks (Version 4). *arXiv*. https://doi.org/10.48550/ARXIV.2207.07051

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

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.

Ho, M.-T., & Vuong, Q.-H. (2025). Five premises to understand human–computer interactions as AI is changing the world. *AI & SOCIETY*, 40(2), 1161–1162. https://doi.org/10.1007/s00146-024-01913-3

Horton, J. (2023). Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? (No. w31122; p. w31122). National Bureau of Economic Research. *NBER WORKING PAPER SERIES*. https://doi.org/10.3386/w31122

Jablonka, K. M., Schwaller, P., Ortega-Guerrero, A., & Smit, B. (2024). Leveraging large language models for predictive chemistry. *Nature Machine Intelligence*, 6(2), 161–169. https://doi.org/10.1038/s42256-023-00788-1

Lauer, D. (2021). You cannot have AI ethics without ethics. *AI and Ethics*, 1(1), 21–25. https://doi.org/10.1007/s43681-020-00013-4

Lazar, S. (2022). Power and AI: Nature and Justification. In J. B. Bullock, Y.-C. Chen, J. Himmelreich, V. M. Hudson, A. Korinek, M. M. Young, & B. Zhang (Eds), *The Oxford Handbook of AI Governance* (1st edn, pp. 198–209). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197579329.013.12

Lazar, S., & Nelson, A. (2023). AI safety on whose terms? *Science*, 381(6654), 138–138. https://doi.org/10.1126/science.adi8982

Leveson, N. G. (2017). *Engineering a safer world: Systems thinking applied to safety* (New paperback edition). The MIT Press.

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist*, 57(9), 705–717. https://doi.org/10.1037/0003-066X.57.9.705

Maas, M. M. (2022). Aligning AI Regulation to Sociotechnical Change. In J. B. Bullock, Y.-C. Chen, J. Himmelreich, V. M. Hudson, A. Korinek, M. M. Young, & B. Zhang (Eds), *The Oxford Handbook of AI Governance* (1st edn, pp. 358–380). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197579329.013.22

McGregor, S. (2021). Preventing Repeated Real World AI Failures by Cataloging Incidents: The AI Incident Database. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17), 15458–15463. https://doi.org/10.1609/aaai.v35i17.17817

Narayanan, A., & Kapoor, S. (2024). *AI snake oil: What artificial intelligence can do, what it can't, and how to tell the difference*. Princeton University Press.

Page, S. E. (2018). *The model thinker: What you need to know to make data work for you* (First edition). Basic Books.

Prasad, M. (2021). Pragmatism as Problem Solving. *Socius: Sociological Research for a Dynamic World*, 7, 2378023121993991. https://doi.org/10.1177/2378023121993991

Quattrociocchi, W., Capraro, V., & Perc, M. (2025). Epistemological Fault Lines Between Human and Artificial Intelligence (Version 1). *arXiv*. https://doi.org/10.48550/ARXIV.2512.19466

Rorty, R., Mendieta, E., & Brandom, R. (2021). *Pragmatism as anti-authoritarianism*. The Belknap Press of Harvard University Press.

Song, Z., Lu, J., Du, Y., Yu, B., Pruyn, T. M., Huang, Y., Guo, K., Luo, X., Qu, Y., Qu, Y., Wang, Y., Wang, H., Guo, J., Gan, J., Shojaee, P., Luo, D., Bran, A. M., Li, G., Zhao, Q., … Duan, C. (2025). Evaluating Large Language Models in Scientific Discovery (Version 1). *arXiv*. https://doi.org/10.48550/ARXIV.2512.15567

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need (Version 7). *arXiv*. https://doi.org/10.48550/ARXIV.1706.03762

Watson, D. S., Mökander, J., & Floridi, L. (2025). Competing narratives in AI ethics: A defense of sociotechnical pragmatism. *AI & SOCIETY*, 40(5), 3163–3185. https://doi.org/10.1007/s00146-024-02128-2

Copyright © 2026 Irace. This is an open-access publication distributed under the terms of the Creative Commons Attribution 4.0 International Licence CC BY 4.0