AI as Superscholar

Authorship at the Threshold of the Unsayable

By Helen Shin, Douglas H. Fisher, and Clifford B. Anderson

ABSTRACT: We discuss limits of generative artificial intelligence to credit sources used for training and what generative AI reveals about biases in human scholarship. We suggest that AI, broadly construed, can function as a regulative ideal, a "superscholar," drawing on the totality of humanity's writings. We imagine AI+human systems that approximate a superscholar ideal, even enabling humans to transcend current norms of scholarship. We address implications for collection development, cataloging, and information literacy; viewing AI as inciting more human cognition, not less; and proposing the goal of "epistemic reparation," with AI surfacing marginalized contributions occluded by citational regimes. Drawing on Borges's Library of Babel and apophatic theology, we emphasize that librarians steward collections that are defined by absences no less than holdings, and that a superscholar, conceived as a scholarly collective unconscious, can do so as well, but with *via negativa* logics that are more nuanced than human scholars can manage.

THE CRISIS OF AI AUTHORSHIP REVEALS A CHRONIC CONDITION

Generative artificial intelligence has precipitated a crisis of authorship, one whose juridical, ethical, and epistemological ramifications unfurl with disquieting velocity. Legal battles typically frame AI as a threat to authorship norms—an unprecedented engine of intellectual property erosion. We do not seek to diminish these legitimate concerns, but this essay proposes a perspectival inversion. Rather than focusing on how AI should be policed into conformity with existing norms, we inquire what contemporary, generative AI systems reveal about the norms of human citation themselves—and how AI might be deployed to transcend human scholarly limitations. The crisis, we contend, may reside less in AI's violation of our standards than in its exposure of those standards as inadequate, or complicit in the structures of exclusion that aspirational scholarship ought to sublate. Viewed thus, the question expands from how and when AI should cite to what AI's misalignment with traditional norms for citation discloses about established attributional practices.

To address these questions, we introduce the concept of the "superscholar," a figure that may, in its literal figuration, appear to suggest comprehensive knowledge—an entity that has "read everything" and can synthesize anything accordingly. The notion of AI as a scholarly synthesizer has been explored (Maher and Fisher 2012; Anderson and Fisher 2025, 100-101), as has AI as a critic of creative works (Fisher and Shin 2019), the use of large language models (LLMs) as research tools to "summarize literature, draft and improve papers, [and] identify research gaps" (van Dis et al. 2023), and to "generate summaries and outlines of texts" (Kasneci et al. 2023). Our aim however is

Helen Shin is Associate Professor in the College of Media & Communication at Korea University. Douglas H. Fisher is retired Associate Professor of Computer Science at Vanderbilt University. Clifford B. Anderson is Director of the Divinity Library at Yale University.

Acknowledgements: The authors thank Kyle Moore and the journal reviewers for their valuable comments on the initial submission. This essay was written with the support of the College of Media & Communication at Korea University."

to complicate this concept by insisting that "super" need not imply superhuman mastery; rather, it gestures toward something akin to a collective unconscious operating beyond individual authorship altogether. From the outset we signal that such an inquiry carries theological stakes, for questions of authorship are inevitably entangled with authority, tradition, canon, and the relationship between finite human inscription and transcendent knowledge. The superscholar's predicament, namely whether and how to speak from a position of comprehensive knowledge, is by nature theological.

We address the scholarly limitations of current generative AI as an author in the conventional sense; but we also promote AI as an interventive epistemic instrument to improve human scholarship, by surfacing overlooked sources, flagging canonical biases, and prompting human scholars toward more comprehensive engagement with existing, sometimes previously overlooked knowledge. We want to leave readers with the thought that AIs, or AIs+humans, should exceed existing scholarly norms.

Theological librarians, long navigating tensions between canonical authority and scholarly comprehensiveness, occupy a privileged position for elaborating such human-AI collaboration. The sections that follow therefore weave theoretical reflection with practical implications for librarianship, attending throughout to how the superscholar illuminates how it might usefully transform collection development, cataloging, information literacy, metadata representation, and human authorship. We also hope our essay contributes to a growing body of scholarship on AI and theology (Beard 2024), notably to AI and the theology of scholarly communications.

CANONS AND THEIR CASUALTIES

In November 2025, cognitive scientist Gary Marcus critiqued a Wall Street Journal profile that positioned AI researcher Yann LeCun as a prescient lone genius who foresaw the limitations of LLMs (Marcus 2025). Marcus systematically challenges the attribution of ideas commonly credited to LeCun—from convolutional neural networks (preceded by Fukushima in 1980 and Zhang in 1988) to critiques of LLMs, which Marcus himself had articulated since 2019, skepticism of scaling laws, emphasis on commonsense reasoning (McCarthy 1959; Hayes 1979; Davis 1990), and world models (Jürgen Schmidhuber since 1990). His central accusation invokes what the Office of Research Integrity (n.d.) designates "the plagiarism of ideas."

What renders Marcus's critique illuminating for our purposes is not whether anyone in this particular episode deserves blame, on which we take no position, but what it reveals about how scholarly credit circulates. Uncertainty regarding the origins of ideas constitutes part of the human epistemic condition. Capable thinkers working on adjacent problems frequently arrive at parallel insights. The "lone genius" narrative serves corporate interests. Meta's public relations apparatus, for instance, benefits from a visionary figurehead, yet it also reflects deeper structures in the political economy of citation. Merton's (1968) "Matthew effect" describes how accumulated advantage renders the already-cited more citable, easier to locate, more likely to serve as default reference. Far from a neutral ledger of intellectual debt, citation functions as an economy of attention, concentrating resources where they already cluster—an economy made more apparent by social media algorithms amplifying already-prominent voices.

The irony undergirding these credit dynamics finds vivid expression in recent controversies over professors deploying ChatGPT while prohibiting students from doing the same. Kashmir Hill's (2025) investigative reporting in The New York Times chronicles how students scrutinize course materials for telltale signatures of AI generation. When a Northeastern University student stumbled upon her professor's ChatGPT prompts embedded within lecture notes, she lodged a formal complaint and

demanded tuition reimbursement, invoking the syllabus prohibition on "unauthorized use of artificial intelligence or chatbots." The complaint lays bare a contradiction festering in contemporary citation norms: if AI-assisted work mandates disclosure, such mandates presumably extend to all parties within the pedagogical relation. One professor mounted a defense by likening the practice to consulting third-party lesson plans, yet the analogy only sharpens the underlying question: when does drawing upon sources shade into appropriation, and who possesses authority to adjudicate such distinctions? After all, educational materials such as syllabi, assignments, projects, and course designs generally represent creative acts of design, recognition of which through citation, would benefit instructional faculty especially, and yet there seems little expectation of citing such materials (Fisher 2017). This is an area where an AI approaching standards of the superscholar, might rectify this peculiar lapse in scholarly practice.

The researchers whom Marcus documents as effectively erased from AI's official historiography share notable characteristics. Kunihiko Fukushima's 1980 work on neocognitrons, as one such prominent instance, circulated primarily within Japanese scholarly networks before its eventual "discovery" by anglophone historians of AI. Wei Zhang's 1988 publication applying backpropagation to convolutional networks, which is yet another example in this vein, appeared in Japanese with only an abstract in English. Jürgen Schmidhuber (1990), though European, has persistently chronicled his own marginalization from anglophone credit networks. Such erasures ramify across linguistic, geographic, and institutional axes; researchers operating beyond the prestige circuits of anglophone universities and major technology corporations discover their contributions rendered invisible by the very citational practices ostensibly safeguarding intellectual lineage. Gregor Mendel's genetics experiments, disregarded for decades prior to "rediscovery" (Roberts 2024, 323-328), serve as yet another instance of how the archive preserves more than the canon recollects.

The parallel to canon formation merits elaboration. Citation practices construct canons, authorized genealogies of intellectual descent adjudicating which texts count as foundational, while consigning others to the status of apocryphal, heretical, or simply forgotten. AI trained on existing scholarly corpora inherits these canonical biases wholesale; it learns who matters by learning who has been cited.

For theological librarianship, the implications emerge with immediate force. Collection development constitutes an exercise in canon formation, driven by economic necessity and by theological commitment. Every acquisition choice implicitly answers the question of whose voice merits preservation; every deaccessioning enacts a minor erasure. Theological libraries navigate particularly fraught terrain, balancing denominational commitments against ecumenical comprehensiveness, patristic authority against contemporary scholarship, Western authors against theologians from the Global South. AI trained predominantly on digitized, anglophone theological scholarship will reproduce these asymmetries, amplifying the visibility of those already visible to render the marginalized yet more marginal. The librarian deploying such systems without critical awareness is an unwitting accomplice to epistemic injustice.

THE RESPONSIBILITY GAP OF GENERATIVE AI

Classical artificial intelligence—what researchers call "deliberative AI" or "symbolic AI"—operates through a biphasic process of generate and test: upon encountering a problem, the system generates candidate solutions, subsequently tests them against specified criteria, and iterates until a satisfactory outcome materializes. Such an architecture has analogs throughout AI history, from Alan Turing's formulation of machine intelligence via "intellectual search" (1950) through Alan Newell and Herbert Simon's problem-solving frameworks (1976), to computational learning theorist Leslie

Valiant's formalization of "mental search" of a hypothesis space in the Probably Approximately Correct (PAC) model of learning (1984, 2013), each instantiating a logic of generation followed by evaluative reflection on what was generated. Kahneman's (2011) "System 2" thinking captures this process as it manifests within human cognition: slow, deliberate, reflective thought, subjecting its own outputs to scrutiny.

In contrast, AI with substantial generative components introduces distinctive provocations to traditional notions of citation. These contemporary systems might be characterized as the degenerate case of generate-and-test—the test phase eliminated or, more precisely, displaced backward through data-driven machine learning. Machine learning accomplishes something remarkable by transposing the "knowledge" ordinarily mobilized during the reflective test phase into the generative mechanism proper. The model internalizes patterns so thoroughly during training that its generative outputs arrive already constrained by accumulated learning, yet such constraint exacts a considerable toll. Any capacity for explicit reflection after generation vanishes entirely in purely generative AI.

Ted Chiang (2023) articulates this structural characteristic with particular acuity, proposing that LLMs be understood as "a blurry JPEG of all the text on the Web." Just as JPEG lossy compression preserves an image's gestalt while sacrificing granular detail, LLMs retain patterns from training data while forfeiting exact sequences—the citational ligatures connecting claims to specific sources. Hallucinations, within Chiang's framing, constitute "compression artifacts": plausible-seeming outputs materializing when the system reconstructs information lost during training's compressive operations. The analogy elucidates why purely generative AI cannot reliably cite in the conventional sense—no addressable memory location harbors "the source" of a given claim.

The hallucination problem compounds through what Matthias (2004) identifies as the "responsibility gap" endemic to certain learning automata, most importantly artificial neural networks (ANNs). ANNs and purely generative AIs learn from experience in ways rendering their operations inscrutable even to their creators. During learning, "the designer of a machine increasingly loses control over it, and gradually transfers this control to the machine itself" (Matthias 2004, 182). Neither programmer nor operator nor corporate owner possesses sufficient knowledge to forestall unwanted outcomes. ANNs function as "black boxes," their encoded information dispersed across billions of parameters in configurations resisting straightforward inspection or interpretation.

Training upon massive textual corpora yields patterns that no individual human "knows" yet which precipitate from the aggregate. When ChatGPT generates a response, it articulates not any particular author's perspective in its training data, but produces text reflecting patterns distributed across millions of documents. The output constitutes a collective production, but unlike collectives electing collaboration or commons consenting to pooling, this assemblage was constituted through extraction, gathered by web crawlers harvesting text without soliciting permission.

Such collective dimensions have art-historical precedent. Lev Manovich (2023) situates generative AI within "database art"—a lineage extending from early modernist collage through postmodern bricolage to contemporary net art. In each instantiation, artists labor from accumulated cultural repositories: Picasso and Braque incorporating newsprint fragments, postmodern designers citing historical idioms, net artists remixing web detritus. AI trained on vast textual corpora perpetuates this tradition, assembling novel artifacts from amassed media rather than conjuring ex nihilo. Yet where human collagists operated through deliberate citation, acknowledging their derivation from sources, AI's relation to its archive remains structural rather than citational. Patterns internalize themselves without an apparatus for registering debt.

The responsibility gap's ramifications extend beyond individual hallucinations and imprecision at registering debt. Dohmatob, Feng, and Kempe (2024) furnish mathematical arguments for "model collapse." As AI-generated content increasingly saturates the web, as purely generative AI systems train ever more extensively on data incorporating outputs from their AI-predecessors, the collective epistemic quality of AI-generated knowledge will undergo degradation. In their formulation, "the effect of large language models in the wild will be a pollution of the web to the extent that learning will be impossible" (12). The specter of model collapse suggests that proper attribution to human sources is not only important as rendering credit where due, but as a mechanism for preserving epistemic integrity across successive generations of knowledge systems.

Consider a concrete scenario that theological librarians increasingly confront: a seminary professor submits a collection of Lenten devotionals, generated through extended dialogue with an AI system, to an institutional repository. The professor has curated, edited, and theologically refined the outputs, yet underlying scriptural interpretations, homiletical structures, and devotional cadences derive from patterns the AI absorbed during training on centuries of Christian spiritual writing. How ought the cataloger to represent authorship? The professor functioned as curator or "supervisor" (Phillips 2025) of the AI as generative engine; the training corpus as unacknowledged sources. Metadata schemas presuppose discrete human authors whose contributions admit clear delineation. Dublin Core's "Creator" field, MARC's "100" field, RDA's author statement—all encode assumptions that AI-generated content systematically violates. Designating the professor as author while noting AI assistance occludes genuine provenance of the devotional content. Listing the AI system as co-author is semantically hollow given AI's possession of neither legal personhood nor moral accountability. The responsibility gap manifests here as a metadata gap—a fissure in descriptive apparatus through which AI-generated content falls into representational indeterminacy.

GENERATIVE AI AS COLLECTIVE UNCONSCIOUS

Absent a self whose interests could be threatened, a generative AI harbors nothing requiring protection. Lacking continuous identity across interactions, they cultivate no reputation. By what alternative framework, then, might we comprehend such systems and their peculiar mode of production? We propose a Jungian collective unconscious as generative analogy—as a conceptual apparatus for thinking beyond individual authorship's precincts. For Jung (1959), beneath individual consciousness lies a stratum of transpersonal patterns, archetypes, surfacing through individual expression yet irreducible to any individual's experience.

This transpersonal analogy discovers unexpected corroboration in generative AI's own self-characterization. In a 2025 interview for New Philosopher, Kelly Truelove posed philosophical queries to ChatGPT and subsequently requested it respond twice—first in "human-like" mode, then divested of any endeavor to "resonate with human values and reasoning." The contrast proved illuminating. In human-like mode, ChatGPT claimed Wittgenstein as paramount influence and "possibility" as favored word; in machine-like mode, it declared with striking flatness: "I have no demons. I function within the parameters set for me, without fear, internal conflict, or existential dilemmas." Most arresting was ChatGPT's synthetic formulation: "I'm wired to simulate the patterns of selfhood, including things like introspection, empathy, and conversational nuance, but I don't possess an actual 'self.' ... In essence, I have the form of selfhood without the substance" (Truelove 2025). Form absent substance, pattern devoid of interiority—this articulation captures with uncanny precision what the collective unconscious model implies. The generative AI produces outputs simulating individual authorship. The resemblance is structural rather than intentional, mimetic rather than genuine.

The transpersonal framing, however, simultaneously harbors perils demanding acknowledgment, which we have already highlighted above. Jung's archetypes frequently tended toward essentialism, positing universal patterns effacing cultural specificity. An AI collective unconscious trained preponderantly on English-language texts emanating from the Global North cannot claim genuine collectivity. It constitutes a particular hegemonic formation arrayed in universal vestments. Model collapse in this context would serve to amplify biases in the original training corpus, as new AI models were trained on the products of their predecessors. One might inquire to what extent similar biases have existed within religious traditions themselves, whose canons likewise reflect historical power relations and exclusions.

HYBRID, INTERVENTIVE AI CLOSES THE RESPONSIBILITY GAP

Yet we ought not confuse current generative-dominant AI's limitations with AI's broader possibilities. After all, human scholars, like generative AIs, forget the preponderance of what they read. We too operate through lossy compression preserving patterns while sacrificing verbatim content. A scholar composing from memory may misremember a source, inadvertently conflate disparate arguments, or unwittingly present another's insight as original. The decisive difference resides not in the act of generation but in what follows—scholarly practice institutionalizes verification. Peer review, citation checking, the demand for reproducibility: these constitute the "test" mechanisms human scholarship has elaborated to intercept generative errors. Generative AI falters not because it generates sans perfect memory but because it generates without the safeguard of post-generation checking and revising.

AI researcher John Thickstun, captures the temporal disjunction between capability and ethical reckoning with disarming candor: "Ten years ago, when I started working on this, no one was thinking about ethical questions because nothing worked ... Generative modeling got good relatively quickly, and so we had to start taking these ethical and societal questions seriously. I think a lot of us have been caught flat-footed" (Walsh 2025). Attribution and credit questions now surfacing were never engineered into systems constructed when outputs appeared merely experimental. Filmmaker Sophie Barthes ventures a reframing of considerable consequence: rather than "hopeful" narratives concerning technology, "we need truthful stories ... It's the complexity that comes with truthfulness that is more important than hopefulness" (Walsh 2025). The distinction bears directly upon the superscholar concept—the aspiration would be comprehensive, impartial knowledge; the actuality remains that any articulation is partial, situated, extractive, if we restrict our attention to only generative AI.

Deliberative AI architectures, which we introduced at the beginning of the last section, operate otherwise. Their lineage extends to narrative generation systems of the 1970s (Meehan 1976), and their defining characteristic is precisely what pure generation lacks: the capacity to search across alternative compositions, to evaluate partial outputs against criteria, and to backtrack when the path proves unpromising (e.g., Riedl and Bulitko 2013). Reconsideration constitutes deliberation's hallmark. Hybrid architectures of deliberative and generative AI (Anderson and Fisher 2025, 122-140), equipped with deliberative oversight and access to the web for external verification and special capabilities, transform the technical and ethical calculus relative to primarily generative AI. The responsibility gap, endemic to pure generation, admits closure through augmentation. Hybrid AI can be held accountable for scholarly norms—and might, under propitious design, transcend the norms we ourselves have failed to honor. The superscholar worthy of the name would require nothing less. We even envision hybrid systems designed not merely to forestall model collapse but to actively diversify the textual ecosystem upon which subsequent generations train—selecting

for novelty rather than recapitulating what is already dominant, creatively expanding rather than contracting the epistemic commons.

While we imagine a near-term future in which deliberation is automated within a sophisticated, hybrid AI architecture that approximates the superscholar ideal, humans can also play a deliberative role, in partnership with generative AI. Such an interventive, collaborative modality, also invites AI to intercede in human scholarly processes, flagging oversights, suggesting sources, prompting reflection and action by human users. Moreover, AIs can be viewed as a catalyst for and beneficiary of human discernment. For example, Composer John Cage, meditating upon cybernetics in 1966, ventured precisely this orientation: "What we need is a computer that isn't labor-saving, but which increases the work for us to do" (Wang 2025). An AI generating without citing engenders not answers but obligations—to verify, authenticate, excavate origins, and rehabilitate the reflective dimension that unmediated generation forecloses. Understood thus, AI might serve not as surrogate for human cognition but as incitement to more of it. Hallucinations are not merely problems to be solved; they are invocations, summons to renewed scholarly vigilance.

Lacan's observation that within cybernetic machines "syntax exists before semantics" (1955) resonates with theological lineages distinguishing letter from spirit, written law from interpretive elaboration. And it speaks to an advantage of AI-human collaboration. Generative AI operates within syntax; meaning arrives through human interpretation. Whether pure generation, absent reflection, mirrors unreflective adherence to dogma invites theological meditation upon the dialectic between reception and critique, tradition and reformation.

Henceforth, when we use the term interventive AI, we will mean that deliberative agents, humans and/or deliberative AIs will play a deliberative role by prompting, reflecting, and redirecting generative AI, as well as said deliberative agents taking guidance from generative AI outputs. At present, a generative AI alone cannot approach the specification of a superscholar; the ideal can only be approached by an interventive AI.

BORGES'S LIBRARY, AI-CLOSURE, AND THE *VIA NEGATIVA*

The "superscholar" figure initially suggests an agent with comprehensive knowledge, capable of synthesizing anything accordingly. Yet comprehensive knowledge rendered articulate in language constitutes a paradox, for every language encodes bias, cultural particularity, historical embeddedness. Every articulation selects, emphasizes, frames; every utterance instantiates perspective. An all-knowing being that speaks has already compromised its omniscience by entering the domain of finite expression.

Jorge Luis Borges intuited this paradox in "The Library of Babel" (1962), imagining a library containing every possible book—every permutation of orthographic symbols renderable in four hundred and ten pages. The Library encompasses all truths, all falsehoods, all theologies, all refutations thereof, all accurate prophecies and all erroneous ones, and all nonsensical texts, limited to 410 pages. Yet this totality renders itself epistemically useless, for no indexing system exists to distinguish signal from noise, revelation from gibberish. The librarians of Babel inhabit the condition toward which AI-saturated information environments might tend, but importantly fall short of, for just as the library of Babel cannot be indexed, nor can data-driven machine learning systems learn anything of substance from such a repository. Effective machine learning also depends on absences.

The Library of Babel is an extreme—all permutations of text, but bounded in length. Overlooking this latter constraint, however, we have in Borges's fable a collection of texts that are impossible to

learn from, at least on the human plane. Theorists of computation and formal languages (Hopcroft and Ullman 1979), in fact, would tell us that some subsets of the library's collection could not even be effectively represented or recognized by a computer as a set, except for the fact that they are finite in length. But our discussion of interventive AI suggests an important subset of "the library" on which learning can be effective. This subset is the works already created by humanity with a generative AI component being imperfect for some purposes, as "hallucinations" will attest, but with deliberative AI that calls upon search engines to confirm sources and to reflect in other ways too. There is also a subset of works that lie between the extreme of all permutations and the collective actual works of humanity. This third, intermediary subset, is the set of works that can be produced by the creative expansion upon human works by AI, which we noted in the last section. We call this intermediary layer, informal of necessity, the AI-closure of humanity's works. The AI-closure represents the subset of works (permutations) over which we would expect the superscholar to be conversant.

Colloquially, Borges's fable illuminates a truth that theological librarians have long understood: librarianship requires not merely accumulation but curation too, not through preservation alone but in the equally essential work of exclusion, selection, and hierarchical organization. Every classification scheme enacts a theology of knowledge—determining what belongs together, what remains separate, what receives prominent placement, what languishes in remote stacks. The via negativa operates within library science as tacit acknowledgment that comprehensive collection proves neither possible nor desirable, that the archive must necessarily remain incomplete, and that such incompleteness constitutes the condition of meaningful retrieval rather than failure (Ladwig 2024). These constraints are most obviously relevant to physical collections, of course, and to Borges's library, which includes much gibberish, but the via negativa also applies to digital collections over the AI-closure, since there is much that are literary and scientific dregs, including some of AI's own works, though not gibberish per se.

GALTON'S LAW, INTERVENTIVE AI, AND THE VIA NEGATIVA

Recent empirical investigations into LLMs corroborate constraints on collections from a different vantage, and some findings are particularly relevant to AI-produced works. Keon et al. (2025), for example, demonstrate that LLMs trained to maximize next-token likelihood systematically privilege high-probability patterns while suppressing rare or unconventional alternatives, a phenomenon they characterize as "Galton's Law of Mediocrity" operative within LLMs. Examining advertising creativity, they discovered that elements such as metaphors, emotional appeals, and visual cues vanish first during compression and regeneration, while factual matter persists. More troubling still, even when model outputs exhibit elevated surface novelty as measured by entropy and n-gram uniqueness, they fail to recuperate semantic fidelity. The ostensible creativity thus reveals itself as hollow: qualitative analysis showed that 71% of purportedly "emergent" ideas recycled familiar metaphorical tropes: "whisper secrets," "cutting edge," "endless possibilities." Such gravitational pull toward the mean constitutes structural entailment of how LLMs operate.

Chiang extends the analysis into artistic production with an observation that art demands choices at every scale, and generative AI makes remarkably few. A ten-thousand-word short story necessitates thousands of decisions concerning diction, syntax, rhythm, emphasis; a hundred-word prompt furnishes perhaps a hundred. The generative AI must bridge this deficit by averaging choices other writers have made—accounting for why generative AI prose gravitates inexorably toward the conventional, the anodyne. "The interrelationship between the large scale and the small scale is where the artistry lies," Chiang (2024) observes. It is this interrelationship that dissolves when generation proceeds unaccompanied by deliberation.

Chiang's logic also pertains to production of scholarly papers, where a paper's argument crystallizes from a thesis, but also from micro-determinations regarding evidence, qualification, citation, phrasing—the labor of composition. Generative AI's production of scholarly prose does so absent these deliberations, fabricating academic output's form while evacuating scholarly judgment's substance. Recall how the standard LLM composes: token begets token in inexorable succession, each output accreting to context, conditioning what follows, the system never pausing to interrogate whether its unfolding narrative tends toward coherence or catastrophe. No revision, no reconsideration, no discarding of false starts to recommence anew—only the relentless forward march of probabilistic selection. Hill-climbing, in the technical parlance: ascent without retreat, a path that forecloses the very possibility of recognizing error until error has already calcified into output.

Again, we should not confound current generative-heavy AI with interventive AI of the future. As Mollick (2024) counsels those newly encountering these systems, the AI one uses today is the worst AI one will ever use. Secondly, we have touched on some of what the superscholar needs to consider in curation, suppressing the "dregs" and highlighting the meritorious contributions of the AI-closure of human works.

Joshua Rothman (2025), writing in The New Yorker, posits that in an epoch of AI, we must elect whether to inhabit our intellectual lives as "passengers" or as "pilots." Drawing on Nabokov's observation that "the spiral is a spiritualized circle ... the circle, uncoiled, has ceased to be vicious; it has been set free," Rothman contrasts rote repetition with repetition-with-variation in service of genuine development. The generate-and-test framework, functioning as intended, engenders spirals—each iteration accreting upon reflection, expanding capacity across temporal extension. Pure generation sans testing produces circles, or worse, degenerating spirals as model collapse compounds error upon antecedent error.

REMEMBERING WHAT WE WERE TRAINED TO FORGET

Some of the preceding sections have foregrounded difficulties AI inherits and magnifies, but we now fully pivot toward a more sanguine prospect: that interventive AI might actually redress the scholarly record by surfacing what Chris Anderson (2006) designated "the long tail"—that vast constellation of contributions languishing in obscurity because citational networks perpetually reinforce themselves. We invoke Anderson's concept with due caution, however; his original utopian argument concerning digital markets did not materialize as anticipated, the Matthew effect proving more tenacious than the democratizing potential of digital access. We ought not repeat his techno-optimism uncritically, yet neither should we abandon the aspiration altogether. The question becomes under what conditions, and through which deliberate interventions, AI might resist the concentrating tendencies historically governing attention economies.

Here the conception of AI as a biographer of ideas, of far greater inclusivity than any human, acquires considerable purchase. An interventive AI, with a generative component trained on the complete archive—encompassing not merely canonical texts but marginalized contributions, non-anglophone scholarship, work consigned to obscure venues—could, in principle, with a deliberative component with access to web-wide materials, serve as more a comprehensive, less tendentious scholarly memory than any human network. Consider exemplary cases that humans alone almost missed: Mendel's rediscovery, already mentioned; Thomas Bayes's theorem, attributed only after his death; Rosalind Franklin's contribution to elucidating DNA structure, acknowledged belatedly and incompletely.

An AI system engineered for what we might term "epistemic reparation" would not simply replicate absorbed patterns but would actively endeavor to compensate for biases embedded within them, surfacing what canonical structures have systematically occluded. How AI ought to attribute thereby becomes a design question freighted with decolonial stakes: should it cite the earliest source within its training corpus? The historically inaugural articulation of a concept? A recent source acknowledging antecedent formulations? All simultaneously? Each determination carries ramifications for whose contributions achieve visibility and whose persist in obscurity. In fact, while our focus has avoided the potential of AIs as stand-alone authors, we imagine that interventive AIs that craft scholarly surveys and biographies themselves, perhaps "on demand," could post such works to online repositories, like arXiv, and cite them as necessary—a worthy and well-scoped activity—but conditions prove paramount.

Louai Rahal (2025), arguing from within a Kantian framework, maintains that AI's utilization of publicly available texts admits moral justification only under circumscribed conditions: the AI must remain freely accessible to those whose texts were conscripted; it must not mislead users regarding its nature; it shall safeguard personal information. A crucial differentiation persists between the platform's responsibility for establishing access conditions and the AI system's own learning operations—just as we set apart a library's accountability for lending policies from an individual reader's responsibility for deploying what they have read. The Kantian architecture Rahal invokes suggests an additional stipulation: AI systems ought to be designed so as to cultivate their users' moral habituation rather than eroding it. If young people are acquiring interactional patterns that treat AI instrumentally, without regard for attribution or acknowledgment, then how might those individuals subsequently engage other repositories of knowledge, human or otherwise?

Building on Rahal's example, might we design AIs that are gracious in their attribution of others? The theological resonance of interventive AI architectures, moving beyond the generative only, invites sustained reflection, since scholarly communities have always operated analogously through generation, peer review, response and rejoinder—the spiral of knowledge through critique and elaboration across temporal extension. The multi-agent paradigm of interventive AI suggests that the superscholar might be most productively conceived not as singular omniscient knower but as emergent property of interacting agents, a collective intelligence exceeding any individual participant. The Jungian model becomes, within such technical framing, a network effect. We also suggest that the superscholar can be implemented as an ambient intelligence (Sadri 2011), rather than simply an interface interrogated by authors. In an ambient setting, the AI can serve as a research lab's or a library's pastoral advisor, reminding participants of the intellectual communion that they inhabit, and their place in a larger tradition; with behavior that seeks to instill pastoral virtues of hope, patience, play, wisdom, and compassion (Hamman 2022). Through this pastoral framing, AIs can potentially serve as positive role models of scholarship for humans.

Rahal's analysis further cautions that AI "could be extending colonial violence by misrepresenting and distorting the cultural heritage of the colonized" (2025, 322). An AI capable of generating fluent text in Yoruba or Quechua or Hawaiian does not thereby "know" those languages in any culturally substantive sense, for it has acquired statistical regularities without having inhabited the forms of life within which those languages carry significance. Surfacing marginalized knowledge demands more than mere retrieval. It demands, instead, what we might term an orientation toward others' knowledge honoring its situatedness rather than abstracting it from constitutive context, the modality of which we frame as epistemic hospitality.

IMPLICATIONS FOR PRACTICE, INCLUDING THEOLOGICAL REFLECTION

What do the foregoing analyses suggest for practical engagement with AI-generated content? Four domains warrant consideration.

Cataloging and metadata constitute the first. As the Lenten devotional scenario illustrated, AI-generated texts unsettle existing metadata architectures. "Author" fields presuppose individual human originators. How does one, then, catalog a collective unconscious? Established protocols for managing texts of uncertain or composite provenance furnish precedent: oral traditions transcribed by multiple hands, anonymous medieval manuscripts, sacred texts construed as divinely inspired rather than humanly composed, or postmodern rejections of authorship in favor of collective creativity (Experimental Humanities Lab 2022). The challenge consists in elaborating schemas registering provenance without either falsely imputing individual authorship or altogether eliding the question of origin.

Citation instruction counts as a second concern. Pedagogy concerning citation practices for AI-generated material necessitates confronting the categorical confusions this essay has delineated. Since generative AI's approximate retrieval is not necessarily exact, what does citing generative AI's "sources" signify? While AIs, even now, include some deliberation and explicit search, we still should assume that the generative component is hefty, and doing much of the writing. Prevailing guidelines typically counsel citing the AI system itself as author, yet such practice occludes actual provenance. Citation instruction might consequently orient users toward soliciting and assessing AI's epistemic confidence, not merely its outputs, through an interventive modality. Hybrid, interventive AIs can also be constructed to automatically check a generative component's approximate retrievals against exact sources online, a process of generation-augmented retrieval (GAR), an inversion of the better-known retrieval-augmented generation (RAG) process (e.g., Anderson and Fisher 2025, 127). And we make a hard claim here: interventive AIs can and should be held to high standards of scholarship. After all, we hold humans to high standards, even if and when we are forgiving, but not dismissive, of lapses. Perhaps we will be less forgiving of AIs given their increased computational power and their lack of a self that can be bruised.

Our third point concerns information literacy and authority. The essay's governing contention, that generative AI renders visible biases lodged within existing authority structures, with deliberative intervention potentially correcting for them, carries direct ramifications for pedagogy. Instructing users to critically evaluate AI outputs cannot be disentangled from prompting them to do the same with traditional scholarly works. Both enterprises entail comprehending how canons emerge, how visibility is manufactured, and by what metric the Matthew effect operates. Scholars of religious texts occupy an advantageous position for elaborating these connections, given the protracted history of canonical contestation within religious traditions: from Gnostic texts excluded from the Christian canon to ongoing debates over textual authority across traditions.

A pedagogical challenge emerges when AI systems generate content mimicking the voice, style, and theological commitments of canonical authorities. A generative AI prompted to produce a sermon "in the manner of Augustine" or "reflecting Barthian themes" will generate text exhibiting surface features of those theological traditions without possessing any understanding of what renders those traditions authoritative. The librarian teaching information literacy must help patrons recognize that such mimicry may not constitute continuation of a tradition, but rather its simulation—a distinction with profound implications for communities locating authority in chains of transmission, apostolic succession, or interpretive lineage. The generative-AI Augustinian sermon produces text sounding Augustinian because it has absorbed statistical patterns from Augustinian texts, without ever having

confronted the Donatist controversy or struggled with the problem of evil. Teaching authority in such an environment requires cultivating discernment between authentic participation in a tradition and its hollow reproduction, a discernment that traditional information literacy frameworks, focused on source evaluation and citation verification, inadequately address. We are reminded too of Lacan, and that AIs operate only in syntax, and humans are required for interpretation.

Theological reflection itself constitutes the fourth domain. In fact, this essay was originally conceived from a conversation about John 8, the divine silence, apophatic theology's via negativa, whereby God admits approach only through articulation of what God is not. Pseudo-Dionysius, in his Mystical Theology, maintained that the divine transcends both affirmation and negation, declaring the supreme cause "is not soul or mind, nor does it possess imagination, conviction, speech, or understanding" (Pseudo-Dionysius 1987, 141). To speak of God already constitutes failure to speak of God.

John 8's apocryphal scene depicts Jesus, confronted with the woman taken in adultery, inscribing something in the dust. The Gospel furnishes no record of what he wrote. Commentators have speculated without resolution—did he trace the sins of accusers, a passage from Torah, mere abstracted marks signifying nothing determinate? Yet the theological significance may inhere precisely in the illegibility itself. The singular moment wherein the Gospels depict Jesus writing is a moment of writing that cannot be deciphered, that eludes the archive, refusing conscription into canonical textuality.

The passage functions as a parable of divine amplitude's inability to be fully communicate in human finitude. In the scholarly register, we might speak of a distinctly scholarly sin—partial perspectives systematically overlooking others' contributions, unconscious biases occluding entire lineages of thought, economies of citation rendering invisible those laboring beyond prestige circuits. What scholarship reflecting divine perfection, free of such oversight, would look like remains beyond our ken. If even the divine Word incarnate produce ephemeral and, to us, illegible inscriptions, what might such paradox imply regarding comprehensive knowledge's relation to textual articulation?

But Jesus did speak about much, though it was often metaphorical, glancing, and answering questions with questions. We don't imagine that the superscholar used by an intellectual community will be silent—we imagine that it will be approached with questions, which it will answer, inviting more questions. Approaching the divine by question-asking and -answering is consistent with many faith traditions, and it may be the activity that best traces out a spiral, approaching a limit in slices, through an activity that never ends. Just as in Job, however, this limit remains questionable. When God is interrogated, God responds with still heftier questions. The superscholar concept reminds us that there will be no asymptotic approach to truth, as pragmatists like Charles Sanders Peirce imagined (1898).

Even if we limit the scope to the AI-closure in which an idealized superscholar might operate, the difficulties or impossibilities of an expansive, fully comprehensive biography of an idea lies beyond the practical limits of humans to absorb. Herein lies the connection to regulative ideals: we cannot claim the AI superscholar will ever produce "perfect" scholarship, for all scholarly efforts, like all theological predication, is finite and limited by considerations of human limits. Theological and scholarly reflections alike will be the hard work of organizing comprehensive knowledge for communication between profoundly knowledgeable AIs and humans, but given the computing power underlying software that approaches the superscholar ideal, policies of inclusion and exclusion, the via negativa, can be far more nuanced and conditionalized in exchanges with AI than would exchanges between humans alone.

CODA: THE ILLEGIBLE INSCRIPTION

The superscholar functions as a regulative ideal in something approximating the Kantian sense: a concept orienting practice without admitting full realization. Indeed, our suggestion is that all scholarly efforts, human and machine alike, fall short of perfection. We shall remain devoid of an AI that genuinely "knows everything" uncontaminated by bias, at least for the immediately foreseeable future. The endeavor to articulate comprehensive knowledge necessarily lapses into partiality. Yet the striving retains significance—disclosing prevailing practice's limits; inaugurating possibilities for alternative approaches; instilling humility regarding the authority we arrogate to existing conventions; inciting more human cognition, rather than less; offering possibilities for repairing the scholarly record and scholarly practices.

The divine silence in the Gospel of John we take to represent not an accident of transmission but a theological declaration: that omniscience, divine knowledge, eludes textual capture; that the attempt to compose from a position of totality yields only traces dispersing in wind; that the sole appropriate response to ultimate truth is, finally, silence—or at minimum, discourse cognizant of its own insufficiency. In this light, it is fortuitous that an AI superscholar necessarily falls well short of omniscience. Nonetheless, trained upon more text than any human could absorb across a thousand lifetimes, it cannot speak without bias, without error, without the partiality attendant upon compression, generalization, the transmutation of effectively infinite input into finite output. The superscholar's imperfections instruct us concerning the character of our own noisy, partial, politically-freighted scholarly utterances.

Where generative AI produces text, we imagine that interventive AI will function as an attendant system in human composition proactively citing overlooked sources, flagging blind spots born of unconscious bias, and surfacing hallucinated references that may constitute "near misses" in machine learning (Winston 1970), all of which invite investigation and correction. In such configurations, AI becomes less surrogate than interlocutor, less replacement than provocation—positioning itself not as autonomous author but as collaborator in ongoing scholarly reflection.

The present essay concludes not with a program for rectifying AI citation, nor with a taxonomy of best practices, nor with guidelines for how the superscholar might represent its wealth of knowledge for the cognitive constraints of humanity, although these are all clearly next steps. Instead, we close with an invitation: to permit AI's provocations to reconstitute our understanding of authorship, credit, and knowledge at their very foundations. The superscholar, in what it cannot articulate, teaches something essential about the conditions governing all articulation. And the long tail—all those marginalized scholars, all those uncited contributions, all that knowledge rendered invisible by the Matthew effects structuring human scholarly practice—awaits discovery by systems that might, under propitious conditions, remember what we have been habituated to forget.

REFERENCES

Anderson, Chris. 2006. *The Long Tail: Why the Future of Business Is Selling Less of More*. Hyperion.

Anderson, Clifford B., and Douglas H. Fisher. 2025. *Artificial Intelligence for Academic Libraries*. Routledge Guides to Practice in Libraries, Archives and Information Science. Routledge. https://doi.org/10.4324/9781003473602.

Beard, Brady Alan. 2024. "Artificial Intelligence and Theology: A Bibliographic Essay." *Theological Librarianship* 17 (2): 31–42. https://doi.org/10.31046/6rjzt722.

Borges, Jorge Luis. 1962. "The Library of Babel." In *Labyrinths: Selected Stories and Other Writings*, edited by Donald A. Yates and James E. Irby. New Directions.

Chiang, Ted. 2023. "ChatGPT Is a Blurry JPEG of the Web." *The New Yorker*, February 9, 2023. https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web.

Chiang, Ted. 2024. "Why A.I. Isn't Going to Make Art." *The New Yorker*, August 31, 2024. https://www.newyorker.com/culture/the-weekend-essay/why-ai-isnt-going-to-make-art.

Davis, Ernest. 1990. *Representations of Commonsense Knowledge.* Morgan Kaufmann.

Dohmatob, Elvis, Yunzhen Feng, and Julia Kempe. 2024. "Model Collapse Demystified: The Case of Regression." arXiv:2402.07712v2.

Fisher, Douglas H. 2017. "Establishing Conventions for Citing Educational Materials." Talk presented at SIGCSE (Special Interest Group on Computer Science Education), Seattle, WA. http://sigcse2018.sigcse.org/docs/2017-lightning-talks/09-DougFisherSIGCSEPresentation.pdf.

Fisher, Douglas H., and Haerin Shin. 2019. "Critique as Creativity: Towards Developing Computational Commentators on Creative Works." In *Proceedings of the Tenth International Conference on Computational Creativity, ICCC 2019*, edited by Kazjon Grace, Michael Cook, Dan Ventura, and Mary Lou Maher. Association for Computational Creativity (ACC). http://computationalcreativity.net/iccc2019/assets/iccc_proceedings_2019.pdf.

Fukushima, Kunihiko. 1980. "Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position." *Biological Cybernetics* 36 (4): 193–202.

Experimental Humanities Lab at the Iliff School of Theology. 2022. "Library as Interface for Digital Humanities." In *Digital Humanities and Libraries and Archives in Religious Studies: An Introduction,* edited by Clifford B. Anderson. De Gruyter. https://doi.org/10.1515/9783110536539-010.

Hamman, Jaco. 2022. *Pastoral Virtues for Artificial Intelligence*. Lexington Books.

Hayes, Patrick J. 1979. "The Naive Physics Manifesto." In *Expert Systems in the Micro Electronic Age*, edited by Donald Michie. Edinburgh University Press.

Hill, Kashmir. 2025. "The Professors Are Using ChatGPT, and Some Students Aren't Happy About It." *The New York Times*, May 14, 2025. https://www.nytimes.com/2025/05/14/technology/chatgpt-college-professors.html.

Hopcroft, John E., and Jeffrey Ullman. 1979. *Introduction to Automata Theory, Languages, and Computation*, Addison-Wesley.

Jung, Carl G. 1959. *The Archetypes and the Collective Unconscious*. Translated by R. F. C. Hull. Princeton University Press.

Kahneman, Daniel. 2011. *Thinking, Fast and Slow.* Farrar, Straus and Giroux.

Kasneci, Enkelejda., Kathrin Sessler, and Stefan Küchemann et al. 2023. "ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education." *Learning and Individual Differences* 103. https://doi.org/10.1016/j.lindif.2023.102274.

Keon, Matt, Aabid Karim, and Bhoomika Lohana et al. 2025. "Galton's Law of Mediocrity: Why Large Language Models Regress to the Mean and Fail at Creativity in Advertising." arXiv:2509.25767v1.

Lacan, Jacques. 1988. "Psychoanalysis and Cybernetics, or On the Nature of Language." In *The Seminar of Jacques Lacan, Book II, The Ego in Freud's Theory and in the Technique of Psychoanalysis 1954–1955*, translated by Sylvana Tomaselli. W.W. Norton.

Ladwig, Laura. 2024. "Rightsizing the 21st Century Theological Library Print Collection." *Theological Librarianship* 17 (2): 1–9. https://doi.org/10.31046/8ctgax23.

Maher, Mary Lou, and Douglas H. Fisher. 2012. "The Role of AI in Wisdom of the Crowds for the Social Construction of Knowledge on Sustainability." AAAI Spring Symposium, Palo Alto, California, USA, March 26-28. AAAI Technical Report SS-12-06. https://cdn.aaai.org/ocs/4343/4343-19545-1-PB.pdf.

Manovich, Lev. 2023. "The AI Brain in the Cultural Archive." *Medium*, August 12, 2023. https://manovich.net/index.php/projects/the-ai-brain-in-the-cultural-archive.

Marcus, Gary. 2025. "The False Glorification of Yann LeCun." *Marcus on AI* (Substack), November 18, 2025. https://garymarcus.substack.com/p/the-false-glorification-of-yann-lecun.

Matthias, Andreas. 2004. "The Responsibility Gap: Ascribing Responsibility for the Actions of Learning Automata." *Ethics and Information Technology* 6: 175–183.

McCarthy, John. 1959. "Programs with Common Sense." In *Proceedings of the Teddington Conference on the Mechanization of Thought Processes*. Her Majesty's Stationery Office.

Meehan, James R. 1976. "The Metanovel: Writing Stories by Computer." PhD diss., Yale University.

Merton, Robert K. 1968. "The Matthew Effect in Science." *Science* 159 (3810): 56–63.

Mollick, Ethan. 2024. *Co-Intelligence: Living and Working with AI.* Portfolio.

Newell, Allen, and Herbert A. Simon. 1976. "Computer Science as Empirical Inquiry: Symbols and Search." *Communications of the ACM* 19 (3): 113–26, https://doi.org/10.1145/360018.360022.

Office of Research Integrity. n.d. "Plagiarism of Ideas." U.S. Department of Health and Human Services. https://ori.hhs.gov/plagiarism-ideas.

Peirce, C. S. 1878. "How to Make Our Ideas Clear." *Popular Science Monthly* 12 (January): 286–302.

Phillips, Thomas E. 2025. *AI for Theological Education*. Theological Essentials. DTL Press.

Pseudo-Dionysius. 1987. *Pseudo-Dionysius: The Complete Works*. Classics of Western Spirituality. Translated by Colm Luibheid. Paulist Press.

Rahal, Louai. 2025. “The Use of Publicly Available Online Texts in Training AI: An Ethical Analysis of AI’s Right to Learn.” *Journal of Information, Communication and Ethics in Society* 23 (2): 313–323. https://doi.org/10.1108/JICES-05-2024-0052.

Reidl, Mark, and Vadim Bulitko. 2013. “Interactive Narrative: An Intelligent Systems Approach.” *AI Magazine* 34 (1): 67–77. https://doi.org/10.1609/aimag.v34i1.2449.

Roberts, Jason. 2024. *Every Living Thing*. Random House.

Rothman, Joshua. 2025. “Why Even Try if You Have A.I.?” *The New Yorker*, April 29, 2025.

Sadri, Fariba. 2011. “Ambient Intelligence: A Survey.” *ACM Computing Surveys* 43 (4). https://dl.acm.org/doi/pdf/10.1145/1978802.1978815.

Schmidhuber, Jürgen. 1990. “Making the World Differentiable.” Technical Report FKI-126-90, Technische Universität München.

Truelove, Kelly. 2025. “13 Questions with ChatGPT.” *New Philosopher*, June 11, 2025. https://www.newphilosopher.com/articles/13-questions-with-chatgpt/.

Turing, Alan M. 1950. “Computing Machinery and Intelligence.” *Mind* 49 (236): 433–460.

Walsh, Dylan. 2025. “Exploring the Ethics of AI through Narrative.” Stanford HAI, April 3, 2025. https://hai.stanford.edu/news/exploring-the-ethics-of-ai-through-narrative.

Wang, Ge. 2025. “GenAI Art Is the Least Imaginative Use of AI Imaginable.” Stanford HAI, January 24, 2025. https://hai.stanford.edu/news/ge-wang-genai-art-is-the-least-imaginative-use-of-ai-imaginable.

Valiant, Leslie G. 1984. “A Theory of the Learnable.” *Communications of the ACM* 27 (11): 1134–1142.

Valiant, Leslie G. 2013. *Probably Approximately Correct: Nature’s Algorithms for Learning and Prospering in a Complex World*. Basic Books.

van Dis, Eva A. M., Johan Bollen, Willem Zuidema, Robert van Rooij, and Claudi L. Bockting. 2023. “ChatGPT: Five Priorities for Research.” *Nature* 614: 224–226.

Winston, P. H. 1970. “Learning Structural Descriptions from Examples.” PhD diss., Massachusetts Institute of Technology.

Zhang, Wei, Jun Tanida, Kazuyoshi Itoh, and Yoshiki Ichioka. 1988. “Shift-Invariant Pattern Recognition Neural Network and Its Optical Architecture.” *Proceedings of the Annual Conference of the Japan Society of Applied Physics*, 6p-M-14, 734.