JEP

Firenze University Press
https://oaj.fupress.net/index.php/jep

OPEN ACCESS

Citation: Granata, P. (2025) Notes on a sociology of generative knowledge: Human-AI epistemic stewardship. *Journal of Emerging Perspectives* 2: 57-71. doi: 10.36253/jep-20882

Received: April 30, 2025

Revised: September 30, 2025

Published: December 20, 2025

© 2025 Author(s). This is an open access, peer-reviewed article published by Firenze University Press (https://www.fupress.com) and distributed, except where otherwise noted, under the terms of the CC BY 4.0 License for content and CC0 1.0 Universal for metadata.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Competing Interests: The Author(s) declare(s) no conflict of interest.

ORCID
PG: 0000-0002-8441-618X

Notes on a sociology of generative knowledge: Human-AI epistemic stewardship

Paolo Granata

University of Toronto, Canada
Email: paolo.granata@utoronto.ca

Abstract. This article develops a sociology of generative knowledge for the age of AI. It treats contemporary systems as hybrid actors within sociotechnical networks and reframes classical anchors – Mannheim's situatedness, Merton's norms, and Latour's distributed agency – for model-mediated inquiry. Generative knowledge is defined as knowledge organized to produce further knowledge through iterative, tool-mediated, and socially embedded processes. On this basis, the paper advances epistemic stewardship as a practical orientation that sustains human agency while harnessing computational acceleration. Stewardship is operationalized through transparency-by-design, provenance and traceability, calibrated trust, independent verification and red teaming, structured challenge routines, inclusivity in data and participation, and proportional delegation to machines. The account clarifies gains in discovery and education alongside risks from opacity, automation bias, feedback loops, and cognitive drift, and it specifies institutional reforms: standardized disclosure artifacts, replication triggers for AI-assisted claims, revised authorship taxonomies, and equitable access to compute, benchmarks, and community-governed datasets. A research agenda follows, calling for comparative evaluations of stewardship designs, longitudinal studies of hybrid practice, and field-specific protocols that link evaluation to adoption thresholds. The result is a framework that integrates social theory with design and governance guidance, aiming to convert acceleration into certified advance while keeping responsibility legible and contestable.

Keywords: generative knowledge, epistemic stewardship, human-AI co-creation, distributed agency, actor-network theory, Mertonian norms, automation bias, situated knowledge, provenance tracking, calibrated trust, communal verification, epistemic literacy.

1. INTRODUCTION

The present moment compresses decades of speculation about artificial intelligence into an everyday epistemic fact: machines now participate, visibly and pervasively, in the making of what we take to be knowledge.

Rather than rehearse the exhausted question of whether machines "think" (Turing 1950), I propose generative knowledge (Granata 2026) as the analytic lens through which to apprehend AI's contribution to inquiry. Distinct from computational epistemology, which centers on the logic of algo-

rithmic inference, and broader knowledge co-production, which emphasizes collaborative social dynamics, generative knowledge defines a specific sociotechnical mechanism. It is the organization of inquiry wherein AI outputs function not as representations of settled fact, but as provisional scaffolds that compel further, iterative human verification and hypothesis formation. In this view, knowledge is defined by its capacity to catalyze a recursive cycle of questions, rather than its status as a static retrieval. From this vantage, AI is neither a mere amplifier of human cognition, or a mere "epistemic enhancer" (Humphreys 2004, p. 6), nor a disembodied oracle (Burrell 2016); it is an actor situated within networks of practice, instrumentation, and norms, helping to open new problem spaces, accelerate hypothesis formation, and reorganize pathways of validation.

The shift matters because the locus of epistemic labor is moving. Large language models, data-intensive pipelines, and agentic interfaces do not simply retrieve or recombine existing material; they instantiate new procedures of search, abduction, and pattern conjecture that reconfigure the division of epistemic functions between humans and machines. If, following Mannheim (1954), knowledge is always situated, and if, following Merton (1942), its public warrant depends upon communal norms and institutions, then the arrival of generative AI systems forces a recalibration of both situatedness and warrant. Where are these systems situated, and by whom? What counts as certified knowledge when inference, suggestion, and synthesis are distributed across human judgment, computational heuristics, and infrastructural stacks? And how do we preserve the virtues of critical scrutiny when fluency and speed threaten to become surrogates for understanding?

This article advances three central claims. First, AI should be approached as a hybrid participant in knowledge networks whose agency is derivative yet consequential: derivative because it remains scaffolded by human purposes, data ecologies, and institutional rules; consequential because its operations systematically shape what is thinkable, searchable, and actionable. Second, the concept of generative knowledge clarifies both the "bright side" and the "dark side" of this hybridity by directing attention to the recursive cycle of question → generation → evaluation and to the conditions under which that cycle yields reliable novelty rather than persuasive error. Third, meeting AI's dual potential requires an affirmative practice I call *epistemic stewardship*: a cultivated stance, at once normative and procedural, that sustains human agency, mandates transparency and traceability, and institutionalizes independent verification where computational capability now expands the frontier of inquiry.

What follows is organized accordingly. After providing a methodological note to clarify scope conditions of the work presented, I first reconstruct the historical and theoretical ground – Mannheim on social situatedness and Merton on the ethos and governance of science (Jasanoff 2004) – to establish a baseline for legitimacy and communal warrant in AI-mediated inquiry. I then recast AI, via Latour's actor-network sensibilities (1993; 2005), as a hybrid actor that redistributes agency, authority, and accountability across humans, models, data, and institutions. With this frame in place, I define generative knowledge and specify its principles, before turning to the anatomy of human-AI co-creation and the calibrated trust it requires. Subsequent sections develop the dual potential: acceleration and access on the one hand; opacity, automation bias, and cognitive drift on the other. I then elaborate epistemic stewardship into principles and consider institutional reforms in science, education, and knowledge economies.

2. METHODOLOGICAL NOTE AND SCOPE CONDITIONS

This article is a work of analytic reconstruction and conceptual synthesis, informed by selective engagement with the classical canon in the sociology of knowledge (Mannheim 1954), the sociology of science (Merton 1942), and science and technology studies (Latour 1993), alongside contemporary scholarship on machine learning (Bender et al. 2021; Bommasani et al. 2021; Nicolae 2025), human-computer interaction (Amershi et al. 2019; Shneiderman 2020), and responsible AI (Burrell 2016; Shneiderman 2020; Crawford 2021; Novelli, Taddeo, and Floridi 2024). The argument develops abductively: it moves between theoretical commitments – Mannheim's social situatedness of knowing, Merton's norm-governed certification of science, and Latour's distributed agency of actor-networks – and observations drawn from recent deployments of generative and predictive systems in research, education, and creative practice (Collins 2025). Sources include peer-reviewed studies of AI-assisted discovery and evaluation, policy documents and standards proposals, and methodological reports on dataset and model documentation.

Several assumptions underwrite the inferences and delimit their scope. The analysis presumes that a) AI systems implicated in knowledge production are not purely oracular black boxes (Burrell 2016; Pasquale 2015) but can be situated within workflows that provide at least partial access to prompts, parameters, datasets, and evaluation metrics (Humphreys 2009); b) that institutions

retain the capacity to enforce disclosure and verification norms (Jasanoff 2003); and c) that human participants possess or can be taught the epistemic virtues (Zagzebski 1996; Baehr 2011) – curiosity, open-mindedness, thoroughness – required for calibrated trust (Lee et al. 2025). The analysis further assumes that the communities in question are oriented toward truth-tracking practices recognizable as communal verification, even where claims are fallible and provisional (Kitto et al. 2023). Under these conditions, the claims advanced here – that AI functions as a hybrid actor in epistemic networks and that epistemic stewardship can convert acceleration into reliable knowledge – are most likely to hold.

The limits of inference are equally important. The argument does not generalize to settings where provenance is structurally inaccessible – highly proprietary platforms such as OpenAI that forbid audit, national security enclaves that block disclosure, or data regimes that preclude any form of external inspection (Pasquale 2015). Nor does it presume transferability to domains where ground truth is intrinsically unsettled and there are no shared standards for adjudicating claims, or to political environments in which institutional independence and adversarial review are compromised (Douglas 2009). Low-resource contexts without adequate compute, connectivity, or linguistic coverage will require adaptations that this article only gestures toward. The present contribution should thus be read as a programmatic map – conceptually thick, empirically suggestive, and normatively oriented – rather than as a census of practice.

3. HISTORICAL AND THEORETICAL GROUNDING

The sociology of knowledge is the sustained inquiry into the social conditions under which knowing becomes possible, credible, and efficacious (Nicolae 2025; Fuchs 2025). It asks how forms of thought arise within patterned relations, how those relations set limits and create opportunities for thinking, and how, in turn, knowledge feeds back into and reorganizes social life. From its classic formulations to its contemporary developments (Babich 2017; Schnettler, Knoblauch, and Raab 2017), the field treats ideas as historically situated, institutionalized, and contested phenomena rather than as free-floating abstractions.

Any attempt to place artificial intelligence within the sociology of knowledge (Collins 2025) must begin, almost inevitably, with Mannheim's insistence that cognition is socially situated (Mannheim 1954; Haraway 1988; Jasanoff 2004). Knowers are not detachable Cartesian points but historically embedded participants whose outlooks are patterned by position, interest, and milieu. For Mannheim, the very possibility of critique – his sociology of knowledge – turns on recovering the conditions under which ideas are formed, circulated, and stabilized. In this light, AI systems inherit rather than transcend situatedness: their training corpora sediment selective histories of discourse (Ferrara 2024); their objective functions encode institutional priorities; their outputs emerge within regimes of use that privilege some audiences, languages, and problems over others. To the extent that AI appears to "hallucinate," to rationalize patterns as if they were realities (Huang et al. 2024), it reenacts – at computational scale – the perennial sociological lesson that representations are never innocent (Ji et al. 2023). The relevant question, then, is not whether AI is biased or epistemically neutral, but how its forms of situatedness are rendered visible, governed, and made answerable to communities of inquiry.

Merton's reconstruction of the scientific ethos provides the complementary half of this evaluative baseline. If Mannheim directs attention to the social conditioning of knowledge, Merton articulates the institutional safeguards through which a public can warrant claims despite, and indeed because of, such conditioning. Universalism, communalism, disinterestedness, and organized skepticism were never mere moral ornaments; they described a set of norms and reward structures that tied credit to disclosure, novelty to replication, and prestige to vulnerability under critique. Under this regime, "legitimate knowledge" is not what a solitary mind believes to be true, but what a community, operating with regulated openness and contestability, can provisionally certify (Humphreys 2009; Kitchin 2014). Citation, authorship, priority rules, and peer review are not ephemera: they are the scaffolding by which facts are stabilized, contributions attributed, and error made corrigible.

When these two traditions are read together, a baseline for evaluating AI's role in knowledge comes into view. Legitimacy requires that the situatedness of AI systems be mapped and governed, while warrant requires that their contributions be rendered commensurable with communal procedures of verification (Collins 2025; Hosseini et al. 2025). Training data must be treated as historically contingent archives; model architectures and prompts as methodological choices; inference traces and uncertainty estimates as disclosures analogous to methods sections. The credit economy must adapt, too (Bauchner and Rivara 2024): if models participate in hypothesis generation, synthesis, or measurement, then authorship and acknowledgment practices must specify which epistemic functions were human-initiated, which were machine-augmented, and how responsibility

was retained for interpretation and error. Absent these adjustments, acceleration risks decoupling from certification, and fluency from reliability.

Institutional norms matter for AI-supported discovery because they convert the sociological facts of production into epistemic virtues. Without disclosure requirements, black-box models substitute authority for argument (Balasubramaniam et al. 2023), without communal access to data provenance and evaluation pipelines, replication devolves into ritual, and without incentives for adversarial testing and independent auditing, automation bias can colonize judgment under the guise of efficiency (Horowitz and Kahn 2024; Alon-Barkat and Busuioc 2023). The lesson is not to resurrect a nostalgic picture of science as a vacuum-sealed republic of letters, but to extend the core Mertonian commitment – organized skepticism under conditions of shared scrutiny – into the computational stack. In practical terms, this means treating dataset curation, model documentation, and uncertainty communication as first-order scholarly practices; it means redesigning review to include methodological probing of machine-mediated steps; it means aligning reward structures with transparency and corrigibility rather than mere throughput.

Thus construed, "legitimate knowledge" in AI-mediated inquiry is whatever survives the conjoint tests of situatedness-awareness and norm-governed certification. In other words, knowledge whose genealogies are traceable, whose methods are inspectable, whose claims are reproducible, and whose attributions are fair. Mannheim reminds us that every claim arrives with a social history; Merton teaches how such claims can be transformed – through institutions – into publicly warranted findings. Hence, any sociology of generative knowledge adequate to the present must hold these insights together: it must refuse the fiction of disembodied intelligence (Bender et al. 2021) while renewing the procedures by which novelty becomes trustworthy (Nosek et al. 2015).

4. FROM TOOLS TO HYBRID ACTORS

To treat AI as mere instrument, a tool, or a service is to miss the sociological transformation unfolding in the architectures of inquiry. It is rather an interface for knowledge, an active enabler in our knowledge ecosystem (Granata 2026, p. 51). Latour's Actor-Network Theory (Latour 2005) counsels against a priori partitions of subject and object; it asks us instead to follow the associations through which heterogeneous elements – humans, codes, datasets, platforms, journals, funding agencies – compose provisional unities capable of acting. In this register, contemporary AI systems are best understood as hybrid actors (Rammert 2012; Peverini 2024): they neither replace human agency nor simply extend it, but reconfigure it by introducing new mediations, translations, and points of leverage. A model's output is not the voice of an oracle, nor a neutral mirror; it is an inscription produced by an endless chain of operations that binds curators, annotators, engineers, maintainers of compute clusters, benchmark designers, reviewers, and users. When such chains hold, the model "acts" in the sense Latour intends: it alters the state of affairs by mobilizing a network that makes certain moves easier and others harder.

Agency in these networks is thus distributed and composite (Rammert 2008; Callon 1987). The question is not who acts, but how action is made possible and shaped by alignments across the network. For instance, a preprint research paper that credits a model with discovering a candidate mechanism is also crediting the training corpus that tuned its priors, the prompts and hyperparameters that channeled its search, the laboratory protocols that constrained verification, and ultimately the editorial conventions that admitted its claims to public view. Each segment inscribes interests into the next, narrowing or widening the corridor of possible outcomes. Human investigators initiate questions, frame problem spaces, and interpret results; AI components surface patterns at scales and speeds unavailable to unaided cognition; data infrastructures stabilize reference by standardizing formats and ontologies; institutional gatekeepers translate provisional outputs into certified contributions (Bowker and Star 1999). Agency is shared because the capacity to make a difference inheres in the alignments, not in any isolated node.

Authority, on this view, is not a property but an achievement: it accrues to those configurations that reliably transform speculation into findings under regimes of scrutiny (Alvarado 2023). A model acquires epistemic authority (Hauswald and Schmechtig 2025) only insofar as its lineage is documented, its performance is characterized across relevant distributions, its failures are legible, and its outputs are rendered auditable within existing grammars of justification. Benchmarks, ablation studies, uncertainty reports, and external validations are not technical niceties; they are the inscriptions that translate machine suggestions into claims intelligible to communities of practice (Hauswald 2025). The journal's methods section, the lab's notebook, the repository's metadata, or the standardized documentation frameworks for machine learning models together function as the mediating apparatus through which authority becomes durable – contingently, revisably, but with enough stability to guide further work.

Network mapping clarifies accountability because it renders visible the sites where discretion is exercised and where error can be introduced, detected, and repaired (Novelli, Taddeo, and Floridi 2024). If an AI-assisted analysis misleads, the question is not simply whether "the model failed," but which alignment failed: Was the dataset unrepresentative? Were spurious correlates smuggled in by a scraping pipeline? Did prompt templates prime a particular explanation? Did review procedures waive independent replication under the pressure of novelty? By tracing associations, one can allocate responsibility without collapsing into either scapegoating the machine or absolving the human. Accountability becomes a matter of redesigning linkages – tightening provenance capture, instituting counter-prompting rituals, diversifying benchmark suites, staging adversarial audits – so that the network resists drift toward convenience and reasserts its capacity for correction.

Seen from this vantage, the shift from tools to hybrid actors does not anthropomorphize machines; it sociologizes inquiry. It instructs us to replace the mythology of solitary genius, human or artificial, with a careful account of how collectives act and generate new knowledge through technical mediators (Granata 2026). It also provides a practical agenda. To the extent that we can specify the inscriptions required for authority, the alignments that distribute agency well, and the traces that enable responsibility to be shouldered rather than shirked, we move from vague celebration or alarm toward a workable stewardship of human-AI knowledge production.

5. DEFINING GENERATIVE KNOWLEDGE

Generative knowledge names a way of organizing inquiry so it repeatedly produces further, credible knowledge. Knowledge here is not a ledger of certified statements but a poised activity that, once initiated, reliably propagates further inquiry. It is *poietic* rather than merely *mimetic* (Granata 2026, p. 6): oriented to making, re-making, and reorganizing the very problem-spaces it addresses. What renders knowledge "generative" is its capacity to convert a single episode of asking and answering into a widening cycle – new questions, new representations, new instruments, new forms of coordination – such that the horizon of what can be credibly asked and done is enlarged rather than merely filled. In that arrangement, the measure of generativity (p. 28) is not novelty for its own sake but the sustained production of further, credible knowledge – an ever-broadening horizon of inquiry that remains tethered to practice, nourished by shared epistemic virtues, and open to transformation.

Several interlocking principles give generative knowledge its shape (p. 8). First, it is iterative (p. 9): results are not endpoints but scaffolds for subsequent moves, with each cycle reworking premises, representations, and measures. Second, it is instrumentally mediated (p. 9): tools do not merely accelerate operations; they constitute the very kinds of perceptions and inferences that can be made, from microscopes and model organisms to corpora, embeddings, and simulators. Third, it is socially situated (p. 20): claims travel only insofar as they are taken up, challenged, and stabilized within communities whose norms of credit, critique, and care furnish the selective pressure that distinguishes fecund lines of work from dead ends. Overall, generative knowledge is creative in a thick sense (p. 32): it recombines concepts, techniques, and exemplars to yield forms that were not merely latent but unknowable absent the recombination.

When AI participates in the cycle question → generation → evaluation, these principles are rearticulated (p. 208). Question formation becomes a site of co-design: prompts, retrieval templates, and representational choices act as boundary objects through which human aims are translated into machine-searchable intents. Generation becomes a plural activity: models propose patterns, counterfactuals, or design candidates at scale, while humans supply domain constraints, analogies, and value-laden desiderata that steer the search away from spurious convenience toward epistemic salience. Evaluation expands beyond correctness checks to include robustness across distributions, legibility of failure modes, sensitivity to measurement error, and fitness for communal uptake, with independent verification pipelines acting as governors on the accelerating engine. The resulting spiral (Nonaka and Takeuchi 1995) is not a handoff but a trajectory in which responsibilities and affordances are reallocated while the normative load-bearing structures – explanation, critique, replication – remain decisive.

A definition adequate to this context can therefore be stated plainly: generative knowledge is knowledge organized so that its validated outputs function as inputs to further inquiry, under conditions that render the transitions – problem formation, tool use, judgment – auditable and corrigible. It asks investigators to externalize their questions in forms that can be iterated upon, to treat instruments as partners whose biases must be characterized rather than ignored, to cultivate habits of calibrated trust that neither romanticize nor demonize machine suggestions, and to embed evaluation in communities capable of dissent and repair. Under these conditions, AI

becomes a vector of generativity (Granata 2026, p. 208) not because it "thinks," but because it helps arrange the world – conceptually, materially, organizationally – so that thinking can advance with accountability.

6. EPISTEMIC AGENCY AND HUMAN-AI CO-CREATION

Epistemic agency in human-AI inquiry is not a monolith but an interplay of initiating, representing, judging, and trusting (Vaccaro 2024; Wu 2025; Lee 2025). The starting impulse – the selection and framing of questions – ought to remain a human prerogative, because it is here that values, purposes, and world-making commitments are set. Questions are never innocent; they delimit what will count as a relevant signal, which outcomes will be celebrated as discovery, and which risks will be tolerated in pursuit of it. Closely allied to initiation is the work of representation: deciding which vocabularies, data ontologies, and modeling choices are apt for a domain. Machines can learn representations; they cannot, unaided, choose what is worth representing or which abstractions are licit given ethical and disciplinary constraints. Finally, evaluative sovereignty must be retained by human communities. Setting criteria of adequacy, weighting kinds of error, and adjudicating tradeoffs among accuracy, interpretability, and social consequence are normative acts that derive their legitimacy from communal deliberation and global ethical frameworks rather than computational performance (Jobin, Ienca, and Vayena 2019).

Within these guardrails, AI can augment three families of epistemic function. First, it can extend perception by widening the search space, surfacing weak regularities, and generating counterfactuals that would be prohibitively expensive to enumerate by hand (Granata 2026, p. 211). This is cognitive extension in the literal sense: new vistas of pattern and possibility become visible through tool-mediated attention (Alvarado 2023). Second, it can shoulder forms of cognitive offloading that free scarce human capacities for tasks where judgment is decisive: preprocessing unruly corpora, constructing baselines, maintaining reproducible pipelines, and monitoring distributional drift (Gerlich 2025). Offloading here is not abdication but a reallocation of effort toward sense-making and critique. Third, it can act as a generator of candidate inferences – hypotheses, designs, explanations – whose value is realized only insofar as they enter human-governed evaluation loops (Fügener 2022). In each case, augmentation does not erase responsibility; it rearranges it by making the handoffs explicit and accountable.

Trust, therefore, must be cultivated rather than presumed. Above all, calibration is the operative virtue (Lee et al. 2025). It asks that confidence track evidence about model provenance, training distributions, and historical performance under conditions similar to those at hand. Calibrated trust is sustained by habits and infrastructures: keeping genealogies of data and models so that lineage is auditable; exposing uncertainty in forms that invite scrutiny rather than conceal it; staging adversarial tests and ablations that reveal failure modes before deployment; and maintaining independent verification channels that can say "no" when the allure of speed tempts a premature "yes." (Koivisto, Pekkarinen, and Hiekkanen 2023). Crucially, calibrated trust is dynamic (Kitto et al. 2023), revised upward when a system proves reliable under variation and downward when it surprises us in ways that are epistemically or ethically costly. The aim is not to find a once-for-all level of faith in machines but to align reliance with the moving boundary of demonstrated competence (Ferrario 2024).

A typology of roles clarifies how these capacities combine in practice. At times the system functions as instrument, akin to a microscope, making latent structure inspectable while leaving interpretation to investigators (Alvarado 2023); here the human is observer-critic, and the locus of authorship remains squarely with the community that sets up, sanity-checks, and situates the readout. In other episodes, such as large language model interactions, the system acts as explorer, proposing traversals of a problem space – parameter sweeps, synthetic exemplars, or design blueprints – that humans winnow and steer; the human becomes navigator-curator, allocating attention and enforcing domain constraints in human-AI collaboration (Chiang et al. 2024). Elsewhere the system plays the role of interlocutor, a conversational foil that helps externalize tacit assumptions, elicit counterexamples, or pressure-test arguments; the human is theorist-editor, using dialogue to refine claims, where the whole process is driven by epistemic vigilance (Sperber et al. 2010). Finally, there are moments when the system serves as sentinel, continuously auditing pipelines for drift, leakage, or reproducibility regressions; here the human is steward-adjudicator, deciding when the alarms justify intervention. Across these role configurations, the distribution of agency is patterned rather than symmetrical, and epistemic authority accrues where initiation, representation, and evaluation are anchored in accountable human practices (Hauswald 2025).

What follows is an epistemic vigilance and trust commensurate with a Generative Critical Thinking Framework (Granata 2026, p. 109). Humans retain the agenda-setting and value-laden decisions that define

what knowledge is for; machines extend reach, compress routine, and multiply candidate moves; communities enforce the norms that convert speed into reliability. When these elements are aligned, co-creation does not dilute human agency; it deepens it by placing judgment where it matters most and by making the conditions of trustworthy reliance explicit, inspectable, and corrigible.

7. THE "BRIGHT SIDE"

The promise of generative systems is easiest to see where inquiry is bottlenecked by scale. In data-rich sciences, models now canvass search spaces that once exceeded any plausible human enumeration, surfacing weak and multivariate regularities from genomic variation (Attwood et al. 2022) to materials phase diagrams (Merchant et al. 2023), to domains across diverse disciplines. Hypothesis generation shifts from artisanal conjecture to systematic proposal: ensembles produce ranked candidates, counterfactuals are synthesized to probe causal structure, and parameter regimes once ignored for lack of bandwidth become experimentally thinkable (Zhang et al. 2025). In design fields, iterative co-creation shortens development cycles from weeks to hours. Alternative architectures, molecular scaffolds, or policy scenarios are produced according to defined specifications, then improved through domain-specific constraints and independent testing (Canty 2025). The acceleration increases the pace of discovery, widening the range of questions that can be explored and lowering the cost of failure.

Education displays a parallel dynamic. For novices, AI serves as a supportive structure that avoids both oversimplification and mechanical repetition. It can restate ideas at suitable levels of abstraction, present exercises that increase in difficulty, and provide timely examples while holding back full solutions until the learner has shown intermediate reasoning. For advanced learners, generative systems become dialectical partners, pressing for counterexamples, suggesting literatures adjacent to a topic, and exposing tacit assumptions through challenge prompts. The result is an enlargement of participation (Kestin 2025): students with uneven preparation can enter disciplinary discourse sooner; non-dominant languages gain more immediate access to resources; small institutions without extensive libraries or laboratories can simulate exploratory analysis and prototype research questions that would otherwise be deferred.

These benefits are not automatic. Turning rapid experimentation into dependable knowledge relies on institutional systems that maintain the friction needed for verification. High-throughput discovery calls for pre-registered analytic plans, clear distinctions between exploratory and confirmatory work, and traceable links among datasets, prompts, models, and outputs within a provenance graph. Fast hypothesis generation must be supported by independent replication, through benchmarks, blinded assessments, and adversarial review to reduce the risk of compelling but false results (Yilmaz and Yilmaz 2023). Educational gains likewise require curricular redesign: assessment that rewards process over product, disclosure norms for AI assistance, and rubrics that require students to critique model outputs, reproduce results, and trace claims to sources (Urmeneta and Romero 2024).

Access gains also turn on governance of the computational commons. If only well-funded centers can train or even run state-of-the-art models, acceleration will exacerbate inequities rather than ameliorate them. Shared model repositories with documented behavior, community-maintained datasets with transparent sampling frames, and compute credits earmarked for under-resourced institutions are preconditions for genuinely widened participation (Hosseini et al. 2025). Where such conditions hold, expert communities can increase throughput without sacrificing standards: review cycles shift from policing trivialities to interrogating claims that matter; novices become contributors rather than consumers; and the ecology of inquiry absorbs speed not as a threat to judgment but as a multiplier of occasions for it.

8. THE "DARK SIDE"

However, acceleration without commensurate scrutiny engenders its own pathologies. The first is opacity, not merely as a property of model internals but as a systemic characteristic of human-machine assemblages (Yampolskiy 2024). With large language models, in particular, exact reproduction is fragile: differences in model versions, and small changes to prompts or context can all shift the output, even when the same code is used. Provenance is often weak as well: training mixes sources with unclear sampling, fine-tuning sets may be proprietary or undocumented, and many preprocessing steps go unlogged, which obscures lineage and hinders audit (González Barman et al. 2024). Under these conditions, the verification labor expected by Mertonian norms is displaced from adjudicating claims to reconstructing pipelines.

A second risk is complacency, sycophancy, and automation bias (Parasuraman and Manzey 2010; Kahn at al. 2024) that leads to the displacement of judgment by default acceptance of machine outputs (Alon-Barkat and

Busuioc 2023; Loru 2025). Authority leaks from people to interfaces: the affordances that make systems usable – autocomplete, ranked suggestions, confidence phrasing – also create a presumption that what is salient is what is true. The problem is amplified when models speak with the idiom of expertise, provide references that look canonical, or emit calibrated probabilities without calibrated uncertainty (Zhai et al. 2024; Loru 2025). In practice, busy researchers carry forward pre-filled code, pre-specified model configurations, or suggested citations because the marginal cost of interrogation is high and the institutional incentive is throughput. Over time, the adversarial posture essential to critical inquiry erodes (Parasuraman and Manzey 2010), and with it the willingness to hold open alternative frames, to search for disconfirming evidence, or to dwell in uncertainty long enough for a better question to form.

A third danger arises from feedback loops that instantiate and then amplify their own errors. Once model-produced text, images, or analyses circulate in the wild, they are scraped back into training corpora, creating recursively self-referential gradients that privilege what the model already says over what the world newly reveals (Shumailov et al. 2024). Recommendation systems trained on engagement metrics funnel attention toward outputs that are legible to models rather than informative to humans; citation engines amplify papers that fit learned templates; preprint ecosystems adopt stylistic norms that maximize model summarizability, narrowing the space of expression and, with it, the space of conjecture. In research evaluation, proxy targets – benchmarks, leaderboards, impact indicators – become goals, triggering distortions that look like progress while decoupling from the underlying phenomena (Thomas and Uminsky 2022). Verification is compromised not by a single error but by a slowly shifting baseline that makes error harder to notice.

The fourth risk is cognitive drift and deskilling. Offloading search, synthesis, and even experimental design increases short-run productivity but may attenuate the development of expert faculties that are constitutive, not ancillary, to knowledge-making: the sense of what is interesting, the practiced feel for signal amid noise, the capacity to hold contradictory models in suspension, the cultivated memory of canonical failures (Gerlich 2025). If novices learn to write, model, or interpret primarily by curating and post-editing machine proposals, they may never internalize the generative heuristics that make independent inquiry possible. The long arc of expertise depends on formative frictions – failed starts, dead ends, painstaking replications – that AI can mask or route around. A community that systematically outsources those frictions risks a thinning of judgment that will only become visible when models err on cases that look ordinary but are not.

Finally, authorship and accountability are scrambled. Hybrid writing obscures lines of intellectual contribution; image and code generation confound provenance; "model-assisted" results invite ghost authorship by systems that cannot bear responsibility, while human contributors who specified prompts, curated failures, or designed evaluation harnesses struggle to claim credit in legacy taxonomies of contribution (Lindebaum and Fleming 2024). Absent strong disclosure norms, traceable artifacts, and enforceable audit rights, the communal capacity to attribute, to challenge, and to sanction is weakened, and with it the culture of accountability on which durable knowledge depends (Resnik and Hosseini 2025). Peer reviewers cannot tell whether a derivation reflects the author's understanding or the model's fluency; editors must decide whether to accept papers whose central insight was algorithmically proposed; institutions face liability when harms flow from workflows in which agency is genuinely distributed (Bauchner and Rivara 2024).

These risks do not amount to a countermarch to pre-AI inquiry (Wu 2025). They indicate specific failure modes by which opaque assemblages, authority by interface, self-referential feedback, and skill atrophy can compromise verification regimes, corrode peer review, and attenuate the slow cultivation of expert judgment.

9. EPISTEMIC STEWARDSHIP

Epistemic stewardship names a practical orientation to inquiry under conditions of hybrid agency (Granata 2026, p. 224). It treats human-AI collaborations not as substitutions but as negotiated partnerships whose legitimacy depends on the cultivation of norms, interfaces, and routines that make reasons visible, challenge thinkable, and authority revisable. The steward's task is neither romantic defense of pre-digital craft nor uncritical accelerationism; it is the design and maintenance of procedures by which knowledge-making can be made trustworthy, revisitable, and widely shareable. What follows is a set of principles, each paired with concrete scholarly practices, that renders this stance actionable.

Transparency-by-design requires that contributions to a claim be inspectable at the level of data, model, prompt, and post-processing. In practice this means disclosure norms that travel with the work: datasets that document provenance, scope, and known failure modes; prompt ledgers and parameter manifests that make generative steps reproducible; uncertainty and confidence

reporting attached to outputs rather than relegated to appendices; and authorship statements that specify where and how AI assistance shaped the result (Resnik and Hosseini 2025). Transparency here is not a metaphysical demand to "open the black box" so much as a pragmatic insistence that enough of the box be externalized to sustain critique and repair (Novelli, Taddeo, and Floridi 2024; Resnik and Hosseini 2025).

Calibrated trust (Kitto et al. 2023; Lee at al. 2025) is the disposition to align confidence with evidential strength and model competence, rather than with fluency or authority cues. It is cultivated through practices that couple machine suggestion to human justification: interfaces that require reason-giving before adoption; confidence elicitation from both humans and models, displayed side-by-side; competence profiles for models that bound their use to validated domains; and "break-glass" controls that route high-stakes decisions to slower, more expert pathways. Over time, calibrated trust is reinforced by feedback: performance dashboards that close the loop between recommendations, human overrides, and downstream outcomes.

Independent verification operationalizes Mertonian skepticism in a hybrid milieu (Alon-Barkat and Busuioc 2023). It institutionalizes second looks that are methodologically disjoint from the original pipeline: adversarial review that reconstructs the analysis with alternative models and perturbations; replication triggers that automatically flag surprising effect sizes, distributional shifts, or heavy reliance on unvetted generators; preregistration for confirmatory phases following exploratory, AI-assisted discovery; and periodic external audits of training data governance and evaluation suites (Bauchner and Rivara 2024). The aim is not to slow inquiry for its own sake but to separate acceleration from credence, ensuring that speed feeds hypothesis generation while certification remains a communal achievement.

The habit of challenge installs dissent as a design feature rather than a social afterthought. Stewards stage structured contestation through red-team/blue-team exchanges, devil's-advocate prompts that force models to surface counterevidence, and "challenge mode" interfaces that default to presenting rival explanations or counterfactuals before convergence is allowed (Thomas and Uminsky 2022). In classrooms and labs alike, the habit is reinforced through graded requirements to refute at least one high-fluency suggestion with literatures or pilot data, and through incentives that recognize error-finding and restraint – negative results, retractions, and corrections – as contributions to the knowledge commons rather than reputational blemishes (Ferrara 2024; Huang et al. 2024).

Traceability secures the chain of custody for claims, connecting a statement to the materials and transformations that made it possible. Practically, this entails provenance graphs embedded in manuscripts and data portals; cryptographic signatures and time-stamps for model versions and training sets; audit logs that record human edits, prompts, and post hoc adjustments; and interoperable identifiers for entities, processes, and environments that enable cross-lab stitching of evidence. Traceability does not presume permanence – it anticipates revision – so it pairs with versioning practices that let communities compare, roll back, and annotate changes across the lifecycle of a claim (Novelli, Taddeo, and Floridi 2024; Shumailov et al. 2024).

Inclusivity expands whose experience and expertise count in the construction and evaluation of knowledge (Haraway 1988). It is enacted by diversifying training corpora with documented consent and representation goals; instituting participatory data governance for communities most affected by model deployment; adopting credit taxonomies that name varied forms of intellectual labor, including curation, annotation, and red teaming; and lowering the computational and licensing barriers that currently gate access to powerful models and datasets. Inclusivity, on this view, is not an extrinsic virtue but an epistemic necessity: more heterogeneous inputs and evaluators yield more robust generativity and fewer blind spots (Jasanoff 2003; Crawford 2021).

Finally, proportionality – the least-automation principle – holds that the level of delegation to machines should be commensurate with model reliability, domain stakes, and the reversibility of error. Proportionality is implemented through tiered workflow checkpoints: low-stakes exploratory work can be freely machine-augmented; high-stakes or irreversible decisions require multi-expert sign-off, slower methods, and independent replication; and any shift upward in automation level demands prior demonstration of competence under conditions that mirror intended use. Proportionality turns stewardship into governance by binding enthusiasm for throughput to a politics of consequence (Ferrario 2024; Horowitz and Kahn 2024).

Taken together, these principles transform stewardship from a slogan into a practical framework. They embed transparency, calibration, verification, challenge, traceability, inclusivity, and proportionality into the everyday mechanics of inquiry – disclosure templates, confidence and competence reporting, adversarial review protocols, replication triggers, provenance tooling, participatory governance, and tiered decision rules. In doing so, they keep generativity tethered to communal norms of certification while leaving intact the speed

and reach that make human-AI partnerships attractive in the first place.

10. REFRAMING THE SOCIOLOGY OF KNOWLEDGE FOR THE AI ERA

A reframed sociology of knowledge begins by returning to Mannheim with new questions. If knowledge is always socially situated, then AI does not dissolve situatedness so much as multiply its sites. Training corpora, benchmark suites, data pipelines, and interface defaults are not inert backgrounds but socially patterned environments that prefigure what can be known, by whom, and under what descriptions (Collins 2025). Mannheim's ideology critique thus expands from the analysis of class and generation to encompass infrastructures, data genealogies, and platform economics. The point is not to replace the diagnosis of social location with a technical determinism, but to show how social location is now laminated through sociotechnical substrates. Ideology survives as curation: choices about inclusion, exclusion, and weighting in model development sediment particular horizons of relevance. Under a generative knowledge regime, unmasking ideology requires provenance work – making visible the lineage of datasets, the histories of annotation, the contours of pretraining corpora – so that situatedness can be inspected rather than presumed (Schnettler, Knoblauch, and Raab 2017).

Merton's problem-space also persists, though its objects have shifted. Communalism, universalism, disinterestedness, and organized skepticism were always institutional achievements, enforced by review, priority adjudication, and replication (Merton 1942). With AI in the loop, these norms do not become obsolete; they become infrastructural requirements. Communalism now demands shareable artifacts – model architectures, data statements, prompt logs – that render hybrid reasoning inspectable. Universalism demands evaluation regimes that travel across domains and demographics rather than metrics tuned to narrow benchmarks (Thomas and Uminsky 2022). Disinterestedness becomes a governance problem at the level of compute access, platform dependencies, and conflict-of-interest disclosures for model providers. Organized skepticism must be retooled as adversarial evaluation, red-teaming, and replication triggers designed explicitly for model-assisted results. What counted as "certified knowledge" under Merton is still certified by a community, but the community must now interrogate models, data, and workflows alongside claims and methods.

Latour's lesson is that agency is distributed across humans and non-humans, and that facts stabilize when networks hold (Latour 1987, 2005). In the AI era, hybrid actors are not only instruments that extend perception; they are co-authors of representational moves, inscribing biases and affordances into the very form of hypotheses, visualizations, and texts. Actor-network terms help us see why accountability becomes a property of configurations rather than of individuals (Novelli, Taddeo, and Floridi 2024). Yet, the expansion of agency does not excuse the abdication of responsibility. Precisely because agency is distributed, authority must be layered and traceable (Wu 2025). Network mapping clarifies who or what set the priors, tuned the parameters, curated the data, selected the prompts, and accepted the outputs into the record. Without such mappings, responsibility dissipates into the ether of "the model," and organized skepticism loses its grip. Latour's compositionist impulse (Latour 1993) – to ask what worlds are being assembled – guides the design of stewardship practices that bind together actors, artifacts, and norms into accountable epistemic collectives.

Across these reinterpretations, several elements remain central. Communal verification still anchors credibility; claims acquire standing through exposure to organized critique, independent checks, and norms of disclosure (Merton 1942; Kitcher 1993; Nosek et al. 2015). Priority and credit still matter, but authorship must be disentangled from automation in ways that honor conceptual contribution, experimental craft, and curatorial labor (Resnik and Hosseini 2025). The virtues of inquiry – epistemic curiosity, open-mindedness, and intellectual thoroughness (Baehr 2011; Granata 2026) – are still the drivers of discovery, now exercised through practices of calibration, counter-prompting, and provenance inspection (Lee et al. 2025; Kitto et al. 2023). What requires rethinking is the unit of analysis and the unit of accountability. Knowing is no longer adequately described at the level of the solitary knower or even the bounded laboratory; it is a property of hybrid ensembles whose reliability depends on the design of interfaces, the transparency of data lineages, the legibility of model behavior (Burrell 2016), and the enforceability of audit (Novelli, Taddeo, and Floridi 2024).

Generative knowledge (Granata 2026), construed as knowledge that produces further knowledge through iterative, tool-mediated, and socially embedded processes, serves as the bridge between classical commitments and contemporary conditions. It reframes Mannheim's question – who knows? – as a question about the formation of epistemic ecosystems and the stewardship of their generativity. It recasts Merton's norms as design constraints for platforms, repositories, and review processes that can convert acceleration into certified advance. And

it deploys Latour's analytics to make responsibility tractable within distributed agency (Latour 2005). If ideology critique once aimed to unmask the hidden hand of social interests, its complement now is infrastructure critique (Pasquale 2015): tracing how data, models, and interfaces tilt inquiry. If the sociology of science once policed priority and replication, its complement now is protocol-building: specifying disclosure, audit, and challenge routines that keep hybrid networks truth-tracking (Resnik and Hosseini 2025; Hosseini et al. 2025). The result is less a rupture with the classical frameworks than their rearticulation: a sociology of knowledge oriented to technology-mediated co-creation, committed to communal verification, and organized around the practical craft of epistemic stewardship.

II. CONCLUSION: TOWARD DURABLE HUMAN-AI KNOWLEDGE PARTNERSHIPS

This article has argued that contemporary AI systems participate in knowledge creation not as inert instruments but as hybrid actors within sociotechnical networks (Latour 1987, 2005), and that their *generativity* – understood as the capacity of knowledge to beget further knowledge through iterative, tool-mediated, and socially embedded processes – both amplifies and destabilizes established epistemic practices (Granata 2026). Read with Mannheim (1954), AI extends situatedness into the strata of data pipelines and platform economics; read with Merton (1942), it renders communal norms into infrastructural requirements; read with Latour (1987, 2005), it distributes agency across human and non-human actors whose alignments must be composed, traced, and held accountable. The upshot is not a relinquishing of verification, credit, or responsibility but their reconstruction under conditions of technological co-creation (Collins 2025). Hence the central claim: durable human-AI knowledge partnerships are possible only under a regime of epistemic stewardship that sustains reliability, equity, and human agency while leveraging computational acceleration.

If stewardship is the practical name for this reconstruction, then near-term work begins with disclosure and provenance. Researchers should treat prompts, datasets, model versions, and post-processing steps as first-class research objects, archived with the same care as code and data, and described through standardized artifacts that render hybrid reasoning inspectable (Bommasani et al. 2021; Balasubramaniam et al. 2023). Educators should reorganize curricula around epistemic literacy (Urmeneta and Romero 2024; Wu et al. 2025): students must learn to ask questions that models can meaningfully engage, to read uncertainty visualizations, to counter-prompt and adversarially test outputs, and to document the provenance of co-created work. Journals and funding agencies should adopt author contribution taxonomies that distinguish conceptual labor, curatorial labor, and computational orchestration (Resnik and Hosseini 2025); require confidence reporting and adversarial evaluation for AI-assisted findings; and institute replication triggers when claims rely substantively on model-generated content. Institutions, finally, should invest in equitable access – shared compute, open benchmark suites, community-governed data resources (Hosseini et al. 2025) – so that the capacity to participate in generative knowledge is not rationed by wealth or geography.

The policy horizon is equally concrete. Governance should move beyond generic "AI use" statements toward field-specific protocols that bind Mertonian norms to contemporary pipelines: transparency-by-design requirements for provenance capture (Balasubramaniam et al. 2023); calibrated trust practices (Lee et al. 2025; Kitto et al. 2023) that pair model suggestions with independent checks; red-team reviews embedded in peer-review workflows; and post-publication audit channels capable of version-aware critique as models and datasets evolve (Lee 2025). Credit economies must be recalibrated so that those who assemble and maintain the conditions of generativity – dataset stewards, benchmark curators, maintainer communities – are recognized as co-producers of certified knowledge (Zhang et al. 2025). Equity requires building compute commons and multilingual, culturally representative corpora (Hosseini et al. 2025) governed with community consent, lest the infrastructures that scaffold generative knowledge narrow rather than widen the horizon of inquiry.

A research agenda follows from these prescriptions. We need comparative studies that measure how different stewardship designs – provenance interfaces, uncertainty displays, challenge modes – alter the accuracy, reproducibility, and inclusivity of hybrid inquiry (Chiang et al. 2024; Vaccaro et al. 2024). We need longitudinal ethnographies of laboratories, classrooms, and editorial boards learning to live with models, tracing how habits of challenge, calibration, and documentation are built or eroded over time (Gao and Wang 2025). We need methodological work on evaluation beyond benchmarks: mixed-methods protocols that tie external validity, demographic robustness, and error characterization to concrete adoption thresholds (Thomas and Uminsky 2022). We need normative analyses of authorship and accountability under distributed agency, and institutional experiments that redistribute credit to curatorial and infrastructural

labor (Resnik and Hosseini 2025). And we need global collaborations to specify culturally plural standards for consent, attribution, and access in the data regimes that feed generative systems (Hosseini et al. 2025).

The promise of generative knowledge is to convert acceleration into advance; its peril is to convert plausibility into credulity (Ji et al. 2023; Yampolskiy 2024; Huang et al. 2024; Zhang et al. 2025; Loru et al. 2025; Granata 2026). Between these lies the craft of stewardship: designing interfaces that invite scrutiny rather than deference (Horowitz and Kahn 2024), building governance that makes responsibility traceable rather than diffuse (Wu et al. 2025), cultivating pedagogies that form judgment rather than outsource it (Gerlich 2025). If Mannheim urges us to surface situatedness, Merton to organize skepticism, and Latour to compose accountable collectives, then the task ahead is to bind these insights into practice – protocols, platforms, and pedagogies that keep human agency at the center of hybrid inquiry. Durable human-AI knowledge partnerships will not emerge by default; they will be made through the daily work of researchers, educators, editors, and policymakers who insist that what is fast also be fair, that what is fluent also be falsifiable, and that what is generative remain, in the deepest sense, genuinely knowledge-making.

REFERENCES

Alon-Barkat, S. & Busuioc, M. (2023). Human-AI Interactions in Public Sector Decision Making: Automation Bias and Selective Adherence to Algorithmic Advice. *Journal of Public Administration Research and Theory*, 33(1), 153–69. https://doi.org/10.1093/jopart/muac007

Alvarado, R. (2023). AI as an Epistemic Technology. *Science and Engineering Ethics* 29(4), 26. https://doi.org/10.1007/s11948-023-00451-3

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., et al. (2019). Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Paper 3. New York: ACM. https://doi.org/10.1145/3290605.3300233

Attwood, S.W., Hill, S.C., Aanensen, D.M., Connor, T.R. & Pybus, O.G. (2022). Phylogenetic and Phylodynamic Approaches to Understanding and Combating the Early SARS-CoV-2 Pandemic. *Nature Reviews Genetics* 23(9), 547–562. https://doi.org/10.1038/s41576-022-00483-8

Babich, B. (2017). The 'New' Sociology of Knowledge. In *Hermeneutic Philosophies of Social Science*. Walter de Gruyter GmbH. https://doi.org/10.1515/9783110528374-013

Baehr, J. (2011). *The Inquiring Mind: On Intellectual Virtues and Virtue Epistemology*. Oxford: Oxford University Press.

Balasubramaniam, N., Lehmkuhl, K., Holmström Olsson, H. & Bosch, J. (2023). Transparency and Explainability of AI Systems: From Ethical Guidelines to Requirements. *Information and Software Technology* 159, 107197. https://doi.org/10.1016/j.infsof.2023.107197

Bauchner, H. & Rivara, F.P. (2024). Use of Artificial Intelligence and the Future of Peer Review. *Health Affairs Scholar* 2(5), qxae058. https://doi.org/10.1093/haschl/qxae058

Bender, E.M., Gebru, T., McMillan-Major, A. & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '21), 610–623. https://doi.org/10.1145/3442188.3445922

Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S. et al. (2021). *On the Opportunities and Risks of Foundation Models*. Stanford Institute for Human-Centered Artificial Intelligence (HAI), Stanford University.

Bowker, G.C. & Leigh Star, S. (1999). *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press.

Burrell, J. (2016). How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms. *Big Data & Society* 3(1), 1–12. https://doi.org/10.1177/2053951715622512

Callon, M. (1987). Society in the Making: The Study of Technology as a Tool for Sociological Analysis. In *The Social Construction of Technological Systems: New Directions in the Sociology and History of Technology*, Wiebe E. Bijker, Thomas P. Hughes & Trevor Pinch (Eds, pp. 83–103). Cambridge, MA: MIT Press.

Canty, R.B. & Abolhasani, M. (2025). Science Acceleration and Accessibility with Self-Driving Labs. *Nature Communications* 16, (3856). https://doi.org/10.1038/s41467-025-59231-1

Chiang, C.-W., Zixuan, L., Zhuoyan, L. & Ming Y. (2024). Evaluating Human-AI Collaboration: A Review and Methodological Framework. *arXiv preprint* arXiv:2407.19098

Collins, H. (2025). Why Artificial Intelligence Needs Sociology of Knowledge. *AI & Society* 40, 1249–1263. https://doi.org/10.1007/s00146-024-01954-8

Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. New Haven, CT: Yale University Press.

Douglas, H. (2009). *Science, Policy, and the Value-Free Ideal*. Pittsburgh: University of Pittsburgh Press.

Ferrara, E. (2024). Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci* 6(1), 3. https://doi.org/10.3390/sci6010003

Ferrario, A. (2024). Justifying Our Credences in the Trustworthiness of AI Systems: A Reliabilistic Approach. *Science and Engineering Ethics* 30(55). https://doi.org/10.1007/s11948-024-00522-z

Fuchs, S. (2025). Classical sociology of knowledge and contemporary constructivism. In *Research Handbook on the Sociology of Knowledge*, F. Collyer (Ed, pp. 192-208). Edward Elgar Publishing. https://doi.org/10.4337/9781800376649.00020

Fügener, A, Grahl, J., Gupta, A. & Ketter, W. (2022). Cognitive Challenges in Human-Artificial Intelligence Collaboration: Investigating the Path Toward Productive Delegation. *Information Systems Research* 33(2), 678–696. https://doi.org/10.1287/isre.2021.1079

Gao, L. & Wang, Y. (2025). Rethinking Laboratory Life in the Age of AI. *AI & Society*, forthcoming. https://doi.org/10.1007/s00146-025-02495-4

Gerlich, M. (2025). AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. *Societies* 15(1), 6. https://doi.org/10.3390/soc15010006

González B.K., Wood, N. & Pawlowski, P. (2024). Beyond Transparency and Explainability: On the Need for Adequate and Contextualized User Guidelines for LLM Use. *Ethics and Information Technology* 26, 47. https://doi.org/10.1007/s10676-024-09778-2

Granata, P. (2026). *Generative Knowledge: Think, Learn, Create with AI*. Hoboken, NJ: John Wiley & Sons.

Haraway, D. (1988). Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies* 14(3), 575–599. https://doi.org/10.2307/3178066

Hauswald, R. & Schmechtig, P. (2025). New Waves in the Philosophy of Epistemic Authority and Expert Testimony: An Introduction to the Special Issue. *Social Epistemology*, May, 1–6. https://doi.org/10.1080/02691728.2025.2496203

Hauswald, R.. (2025). Artificial Epistemic Authorities. *Social Epistemology* 1-10. https://doi.org/10.1080/02691728.2025.2449602

Horowitz, M.C. & Kahn, L. (2024). Bending the Automation Bias Curve: A Study of Human and AI-Based Decision Making in National Security Contexts. *International Studies Quarterly* 68(2), sqae020. https://doi.org/10.1093/isq/sqae020

Hosseini, Mohammad, Bastian Greshake Tzovaras, Koray Karaca, and Tony Ross-Hellauer. (2025). "Open Science at the Generative AI Turn: An Exploratory Analysis of Challenges and Opportunities." Quantitative Science Studies. https://doi.org/10.1162/qss_a_00337

Huang, L. & Ting, L. (2024). A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems* 43(1). https://doi.org/10.1145/3703155

Humphreys, P. (2004). *Extending Ourselves: Computational Science, Empiricism, and Scientific Method*. Oxford: Oxford University Press.

Humphreys, P. (2009). The Philosophical Novelty of Computer Simulation Methods. *Synthese* 169(3), 615–626. https://doi.org/10.1007/s11229-008-9435-2

Jasanoff, S. (Ed) (2004). *States of Knowledge: The Co-Production of Science and Social Order*. London: Routledge.

Jasanoff, S. (2003). Technologies of Humility: Citizen Participation in Governing Science. *Minerva* 41(3), 223–244. https://doi.org/10.1023/A:1025557512320

Ji, Z., Lee, N., Frieske, R. et al. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys* 55(12). https://doi.org/10.1145/3571730

Jobin, A., Ienca, M. & Vayena, E. (2019). The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence* 1, 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kahn, L., Probasco, E.S. & Kinoshita R. (2024). *AI Safety and Automation Bias: The Downside of Human-in-the-Loop*. Washington, DC: Center for Security and Emerging Technology.

Kestin, G., Miller, K., Klales, A., Milbourne, T. & Ponti, G. (2025). AI Tutoring Outperforms In-Class Active Learning. *Scientific Reports* 15, 97652. https://doi.org/10.1038/s41598-025-97652-6

Kitcher, P. (1993). *The Advancement of Science: Science without Legend, Objectivity without Illusions*. Oxford: Oxford University Press.

Kitto, K., Bruza, P., Sheng Loh, B. & Boschetti, F. (2023). A Quantum Model of Trust Calibration in Human-AI Interactions. *Entropy* 25(9), 1362. https://doi.org/10.3390/e25091362

Koivisto, M., Pekkarinen, S. & Hiekkanen, K. (2023). Transparency and Explainability of AI Systems: From Ethical Guidelines to Requirements. *Information and Software Technology* 159, 107202. https://doi.org/10.1016/j.infsof.2023.107202

Latour, B. (1987). *Science in Action: How to Follow Scientists and Engineers through Society*. Cambridge, MA: Harvard University Press.

Latour, B. (1993). *We Have Never Been Modern*. Cambridge, MA: Harvard University Press. Originally published (1991).

Latour, B. (2005). *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford: Oxford University Press.

Lee, D., Pruitt, J., Zhou, T., Du, J. & Odegaard, B. (2025). Metacognitive Sensitivity: The Key to Calibrating Trust and Optimal Decision Making with AI. *PNAS Nexus* 4(5), pgaf133. https://doi.org/10.1093/pnasnexus/pgaf133

Lindebaum, D. & Fleming, P. (2024). ChatGPT Undermines Human Reflexivity, Scientific Responsibility and Responsible Management Research. *British Journal of Management* 35(2), 566–575. https://doi.org/10.1111/1467-8551.12781

Loru, E., Nudo, J., Di Marco, N., Santirocchi, A., Atzeni R., Cinelli, M., Cestari V., Rossi-Arnaud, C. & Quattrociocchi, W. (2025). The Simulation of Judgment in LLMs. *PNAS* 122(42): e2518443122. https://doi.org/10.1073/pnas.2518443122

Mannheim, K. (1954). *Ideology and Utopia: An Introduction to the Sociology of Knowledge*. New York: Harcourt Brace.

Merchant, A., Batzner, S., Schoenholz, S.S., Aykol, M. et al. (2023). "Scaling Deep Learning for Materials Discovery." *Nature* 624 (7990): 80–85. https://doi.org/10.1038/s41586-023-06735-9

Merton, R.K. (1942). The Normative Structure of Science. In *The Sociology of Science: Theoretical and Empirical Investigations*, N.W. Storer (Ed, pp. 267–278). Chicago: University of Chicago Press, (1973).

Nicolae, S.. (2025). The 'New' Sociology of Knowledge and the Sociology of Valuation. In *The Routledge International Handbook of Valuation and Society*, T. Peetz, A.K. Krüger & H. Schäfer (Eds, 1st ed., vol. 1, 79–90). Routledge. https://doi.org/10.4324/9781003229353-9

Nonaka, I. & Takeuchi, H. (1995). *The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation*. New York: Oxford University Press.

Nosek, B.A., Alter, G., Banks, G.C. et al. (2015). Promoting an Open Research Culture. *Science* 348 (6242): 1422–1425. https://doi.org/10.1126/science.aab2374

Novelli, C., Taddeo, M. & Floridi, L. (2024). Accountability in Artificial Intelligence: What It Is and How It Works. *AI & Society* 39(4), 2331–2343. https://doi.org/10.1007/s00146-023-01635-y

Parasuraman, R. & Manzey, D.H. (2010). Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors* 52(3), 381–410. https://doi.org/10.1177/0018720810376055

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms That Control Money and Information*. Cambridge, MA: Harvard University Press.

Peverini, P. (2024). *Bruno Latour in the Semiotic Turn: An Inquiry into the Networks of Meaning*. Berlin: Springer Nature.

Rammert, W. (2012). Distributed Agency and Advanced Technology. Or: How to Analyze Constellations of Collective Inter-Agency. In *Agency without Actors? New Approaches to Collective Action*, J.-H. Passoth, B. Peuker & M. Schillmeier (Eds, pp. 89–112). London: Routledge.

Resnik, D.B. & Hosseini, M. (2025). Disclosing Artificial Intelligence Use in Scientific Research and Publication: When Should Disclosure Be Mandatory, Optional, or Unnecessary? *Accountability in Research*, March, 1–13. https://doi.org/10.1080/08989621.2025.2481949

Schnettler, B., Knoblauch, H. & Raab, J. (2017). The 'New' Sociology of Knowledge. In *Hermeneutic Philosophies of Social Science*, B. Babich (Ed, pp. 237–266). Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110551563-013

Shneiderman, B. (2020). "Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy." *International Journal of Human-Computer Interaction* 36 (6): 495–504. https://doi.org/10.1080/10447318.2020.1741118

Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R. & Gal, Y. (2024). AI Models Collapse When Trained on Recursively Generated Data. *Nature* 631, 755–59. https://doi.org/10.1038/s41586-024-07566-y

Sperber, D., Clément, F., Heintz, C., Mascaro, O., Mercier, H., Origgi, G. & Wilson, D. (2010). Epistemic Vigilance. *Mind & Language* 25(4), 359–393. https://doi.org/10.1111/j.1468-0017.2010.01394.x

Thomas, R.L. & Uminsky, D. (2022). Reliance on Metrics Is a Fundamental Challenge for AI. *Patterns* 3(5), 1–12. https://doi.org/10.1016/j.patter.2022.100476

Turing, A.M. (1950). Computing Machinery and Intelligence. *Mind* 59(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Urmeneta, A. & Romero, M. (Eds) (2024). *Creative Applications of Artificial Intelligence in Education*. Springer Nature.

Vaccaro, M., Almaatouq, A. & Malone, T. (2024). When Combinations of Humans and AI Are Useful: A Systematic Review and Meta-Analysis. *Nature Human Behaviour* 8, 2293–2303. https://doi.org/10.1038/s41562-024-02024-1

Wu, J.-Y, Lee, Y.-H., Sing Chai, C. & Tsai, C.-C. (2025). Strengthening Human Epistemic Agency in the Sym-

biotic Learning Partnership with Generative Artificial Intelligence. *Educational Researcher* 54(6), 358–368. https://doi.org/10.3102/0013189X251333628

Yampolskiy, R.V. (2024). *AI: Unexplainable, Unpredictable, Uncontrollable*. CRC Press.

Yilmaz, R. & Yilmaz, F.G.K. (2023). The effect of generative artificial intelligence (AI)-based tool use on students' computational thinking skills, programming self-efficacy and motivation. *Computers and Education: Artificial Intelligence* 4, 100147. https://doi.org/10.1016/j.caeai.2023.100147

Zagzebski, L.T. (1996). *Virtues of the Mind: An Inquiry into the Nature of Virtue and the Ethical Foundations of Knowledge*. Cambridge: Cambridge University Press.

Zhai, C., Wibowo, S. & Li, L.D. (2024). The Effects of Over-Reliance on AI Dialogue Systems on Students' Cognitive Abilities: A Systematic Review. *Smart Learning Environments* 11(1), 28. https://doi.org/10.1186/s40561-024-00316-7

Zhang, Y. & Zenil, H. (2025). Exploring the Role of Large Language Models in the Scientific Method. *Nature Reviews Methods Primers* 5, 19. https://doi.org/10.1038/s44387-025-00019-5