The End of the Artisanal Paper:

Knowledge Production in the Age of the Hybrid Intelligent Team

Robert G. Eccles and Shiva Rajgopal
Oxford Saïd Business School and Columbia Business School

May 25, 2026

Abstract

Academic knowledge production is undergoing a structural transformation. Artificial intelligence has not merely accelerated scholarship—it has changed its organizational form. The traditional cottage-industry model, one scholar hand-crafting one paper over three to five years, is giving way to the Hybrid Intelligent Team (HIT): a structured configuration of human and AI agents whose outputs emerge from iterative loops rather than isolated minds.

We examine the AI policies of the six leading accounting and finance journals and find every one built on flawed premises—a failure confirmed by the fact that only 0.1 percent of papers since 2023 have disclosed AI use despite widespread adoption. We propose three reforms: unrestricted AI use, a flipped disclosure default requiring authors to declare when they did *not* use AI, and AI-assisted peer review.

Beyond journal policy, we argue that AI shifts the binding constraint from producing knowledge to recognizing it. This reframes what tenure should select for, what journal style requirements actually signal, how PhD programs should train researchers, and what reduced teaching loads must now justify. The scholars, institutions, and journals that thrive will be those that stop obsessing over who wrote the sentence and start asking whether the knowledge is actually true and useful.

The End of the Artisanal Paper:

Knowledge Production in the Age of the Hybrid Intelligent Team

Introduction: The 1926 Workflow

The way we produce knowledge in accounting and management is broken and has been for some time. For decades, we have operated on a cottage-industry model: a lone scholar, or a small team, spends three to five years hand-crafting a single empirical paper. We treat these papers like artisanal watches—delicate, time-consuming, and obsessed with identifying the individual intellectual leader. The resulting output is often unreadable to practitioners, policymakers, and regulators who most need the insights, and arrives years after the question it addresses has moved on.

The introduction of large language models and multi-agent AI systems is not just a new tool for citation management. It is a fundamental shift in the epistemological form of research itself. We are moving from the individual author to what we call the Hybrid Intelligent Team (HIT). A HIT is not just a person with a ChatGPT subscription. It is a structured configuration of human and AI agents with genuinely differentiated roles: one agent generates, another challenges, a third verifies independently without knowing what the others concluded. A human orchestrator holds the whole together—directing the process, evaluating the outputs, and taking responsibility for the claims. The unit of production is not any individual agent. It is the iterative loop connecting them. And if we do not change our institutional rules—tenure requirements, disclosure policies, frequently unreadable journal styles—we are simply using 2026 technology to prop up a 1926 workflow. It is time to stop worrying about who wrote the sentence and start worrying about whether the knowledge is actually useful.

To see how far the current system has drifted from reality, consider a thought experiment. Two accounting professors are having lunch. One mentions an empirical puzzle she noticed in earnings announcements; the other asks a clarifying question; an idea forms. By dessert, there is the rough shape of a paper. Neither professor can say, afterward, who had the idea. The conversation produced it.

Now run the same scenario at 11 pm. The same professor opens a chat window and types a question. AI responds with a clarification and a reframing. She pushes back. AI concedes one point and extends another. By midnight, the rough shape of a paper exists. Again: who had the idea? The honest answer, in both cases, is that we do not know—and it does not really matter. What matters is whether the idea is good. Whether it advances our understanding of how markets work, how firms behave, how capital gets allocated. Whether it is true.

In our own collaboration, we often cannot identify who originated a particular insight. We generate, challenge, discard, and refine. What survives is better than what either of us would have produced alone. AI enters this process in a structurally similar way. The relevant unit of analysis

is not the agent. It is the loop. Once the loop is the unit, authorship, attribution, and disclosure become secondary questions. The primary question is simpler and harder: did this process produce knowledge? This paper takes that question seriously and works through the implications for how knowledge is produced, evaluated, and rewarded.

1. Is AI Your New Colleague or Your Secretary?

The first step is to stop thinking of AI as a glorified spell-checker. Think of it instead as a colleague who can process information at speeds humans cannot touch, who has read almost everything, who does not get tired or defensive, and who will tell you when your argument has a hole in it—even at 2 am.

The deepest barrier to good scholarship is not logistical. It is ideational. Most researchers, most of the time, are not blocked by the literature review or the coding task. They are blocked by not knowing what question is worth asking, not seeing the anomaly in the data that actually matters, not having the conceptual vocabulary to frame what they have found. That is the hard part. No AI fixes it. What AI does is collapse everything that comes after the idea: the literature review that took three months, the coding task that ate six weeks, the second-best paper that got written because the truly ambitious one would have required expertise in a second discipline the researcher did not have.

These were real frictions and removing them is genuinely significant. But they were never the source of the intellectual value. They were the packaging operation. AI is the world's most powerful packaging operation, and it now runs at near-zero cost. What that means is not that scholarship gets easier. It means that ideation—the capacity to identify a question that matters, frame it precisely, and recognize when the answer is actually an answer—becomes the only thing that separates good research from competent noise. The constraint has always been the idea. AI just makes that fact impossible to ignore.

Think of it the way a senior partner at a law firm thinks about junior associates. The senior partner does not draft every brief. She reviews, shapes, pushes back, and makes the final judgment calls that require experience and accountability. The juniors do the initial work; the senior does the decisive work. No one thinks the senior partner deserves less credit for the brief because she did not draft every page.

Loop-Level Ideation

We call this loop-level ideation. When the loop is functioning, knowledge is not produced by the human or the machine—it is produced by the iterative cycle between them. The test is generative consequence: did the output change what you asked next? Did it produce a distinction you could not have made before the loop ran? Did it move the inquiry somewhere neither the human nor the machine was already going? When the answer is yes, that is loop-level ideation. When the answer

is no—when the human had the idea and the machine dressed it up—that is AI-assisted drafting. Both are real. Only one justifies the shift in how we think about attribution and intellectual contribution.

When we work this way, we honestly do not know who created which idea. Frankly, we do not care. We are both better off for the interaction. If the resulting insight helps a CFO understand financial reporting quality or helps a director assess board readiness, why does it matter whether the spark came from a silicon chip or a carbon-based brain?

The loop analogy, however, obscures one important asymmetry. When two humans collaborate, both bear professional stakes in the output. AI has no stake, no career, no professional consequences. It cannot be sanctioned, retracted, or held to account. The loop distributes ideation; it does not distribute responsibility. The human in the loop is not merely the last person to review the output. They are the only person who can be answerable for it.

2. Should We Flip the Disclosure Default?

Current AI disclosure policies in top accounting and finance journals are based on a fallacy. They demand that authors disclose if and how they used AI—a vestige of the attribution obsession, the idea that we must peel back the layers to find the human core. This is increasingly impossible to satisfy honestly, and the attempt to satisfy it creates compliance theater. We propose the exact opposite: the default assumption should be that AI was used. The only disclosure that matters is when an author has not used AI.

Why flip the default? We offer three reasons. First, the epistemological case: requiring authors to specify what AI contributed, as if that is a tractable question, is both practically impossible and conceptually confused. Second, the asymmetry case: the researcher who did not use AI is making a stronger claim—that their work was produced without a tool most of their peers now use—and that is genuinely informative. Third, the incentive case: requiring AI disclosure creates an incentive to hide AI use, especially for junior scholars worried about perception. We caveat that this third argument is the weakest, since the evidence below shows 99.9 percent of authors are already ignoring existing disclosure requirements—which cuts both ways.

We should say clearly: none of this is a license for fraud. If AI generates a result that is wrong and the author presents it as their own carefully verified finding, that is misconduct. The responsibility for verification sits with the human author, unconditionally. Our proposal changes who has to disclose what. It does not change who is accountable.

3. What Do the Top Journals Actually Say—and Why Does Every Policy Miss the Point?

Before arguing for what journal AI policies should look like, we need to be precise about what they currently say. We examined the AI policies of the three leading accounting journals—The Accounting Review (TAR), Journal of Accounting Research (JAR), and Journal of Accounting and Economics (JAE)—and the three leading finance journals—Journal of Finance (JF), Journal of Financial Economics (JFE), and Review of Financial Studies (RFS). The results are, at once, remarkably uniform and strikingly inadequate.

Journal	Publisher	Permitted AI use	Disclosure requirement	Reviewer/editor AI	Our position
The Accounting Review (TAR)	AAA	Any stage—writing, editing, analysis. Grammar/spell-check exempt.	Mandatory AI disclosure statement after abstract, published with article.	Same disclosure norms apply to editors and reviewers.	✗ Disclosure default wrong; ✗ scope too narrow
Journal of Accounting Research (JAR)	Wiley / Chicago Booth	Permitted as a 'companion,' not a replacement. Language and readability assistance allowed.	Detailed disclosure in Methods section required. Must review IP terms of AI tool used.	No explicit prohibition, but Wiley's general guidance discourages confidentiality breaches.	✗ 'Companion only' unworkable; ✗ Methods disclosure wrong place
Journal of Accounting and Economics (JAE)	Elsevier	Updated Sept. 2025: permits synthesis, identifying gaps, generating ideas. No AI authorship or images.	Separate declaration: tool name, version, reason for use.	PROHIBITED.	✓ Expanded scope welcome; ✗ reviewer ban counterproductive; ✗ disclosure default backwards
Journal of Finance (JF)	Wiley / AFA	AI permitted to assist in drafting/editing text or improving code. Full prompt documentation required for AI-generated datasets.	Authors may briefly describe AI use; no separate declaration form. Prompt documentation required for AI-generated data.	PERMITTED. AI may assist referees in drafting/editing reports and looking up references.	✓ Permissive stance welcome; ✓ prompt documentation good practice; ✗ disclosure default backwards
Journal of Financial Economics (JFE)	Elsevier	Substantially expanded Sept. 2025: permits synthesis, identifying gaps, generating ideas. No AI-generated images.	Separate declaration required. Template: 'During preparation the author(s) used [TOOL] in order to [REASON].'	PROHIBITED. Cites confidentiality and argues critical assessment is 'outside the scope of this technology.'	✓ Expanded scope welcome; ✗ reviewer ban makes system asymmetric; ✗ disclosure default backwards
Review of Financial Studies (RFS)	OUP / SFS	Simplified: AI assistance permitted; authors take full responsibility.	Minimal: authors affirm responsibility at submission. No separate AI Use Declaration Form.	Not explicitly addressed.	✓ Simplified stance welcome; ✗ disclosure default backwards; reviewer guidance absent

The findings across all six journals reveal four specific dissonances—each reflecting a foundational misunderstanding of what AI is doing to the research process.

Dissonance 1: The disclosure direction is backwards. Every journal requires authors to disclose if they used AI. We propose the exact opposite. The empirical case against the current approach is

already in. A study covering 5,114 journals and over 5.2 million papers found that, despite 70 percent of journals adopting disclosure-focused AI policies, researchers' AI use has increased dramatically—with no significant difference between journals with or without policies. Of the 75,000 papers published since 2023, only 76—approximately 0.1 percent—explicitly disclosed AI use. The current disclosure regime is not producing transparency. It is producing the appearance of a policy while researchers quietly ignore it.

Dissonance 2: The permitted scope of AI use has been narrower than practice. JAR still limits AI to a "companion to the writing process, not a replacement." Consider what that means in practice: an author cannot use AI to help identify relevant literature—that is research, not companionship. She cannot use it to check the internal consistency of an argument—that is content substitution. The policy draws a line that no serious AI user recognizes as real. Its continuing effect is not to limit AI's role in research. It is to ensure that the AI use that does occur goes undisclosed. Elsevier's September 2025 update broadened its permitted scope meaningfully for JAE and JFE—a genuine improvement—but the disclosure default remains backwards across all six journals.

Dissonance 3: The reviewer prohibition makes the system structurally asymmetric. Both JAE and JFE explicitly prohibit reviewers and editors from using AI on submitted manuscripts. One of us is an editor receiving 350 submissions per year—many of them AI-enhanced—and is expected to evaluate each one without recourse to any of the same capabilities that made those submissions cheap to produce. The right answer to the confidentiality concern is institutional AI infrastructure, not a blanket prohibition.

Dissonance 4: The Journal of Finance's treatment of AI use by referees points in the right direction. Of the six journals, JF alone permits AI to assist referees in drafting reports and looking up references. More consequentially, JF's prompt-documentation requirement for AI-generated datasets—requiring open-source or time-stamped proprietary models with full documentation of prompts used—is the only provision in any of the six policies that addresses the verification problem rather than just the disclosure problem. That it appears in the finance journals rather than the accounting journals is itself a comment on whose community has moved faster.

From Diagnosis to Reform: Three Proposals

Proposal 1: Unrestricted AI use. The six journals now occupy a meaningful spectrum on permitted scope, and the trajectory is clearly toward broader permissions. JFE and JAE now explicitly permit AI to synthesize complex literature, identify research gaps, and generate ideas—a substantive expansion we acknowledge as genuine progress. The outlier in the direction of restriction is JAR, which continues to limit AI to a companion to the writing process. That framing is operationally unworkable, and since JAR's policy is driven in part by Wiley publisher-level guidance, the most efficient path to reform is probably to engage Wiley directly rather than the journal editors alone.

Proposal 2: Flipping the disclosure default. All six journals have the disclosure default backwards—all require disclosure if AI was used, none requires disclosure if it was not. The failure is universal, but its severity varies by implementation. TAR's is the most demanding: authors must provide a statement whether or not AI tools were used, published with every article. RFS is the least burdensome. If any of the six journals moved first to a flipped default, our prediction is that adoption would be rapid because the flipped norm removes the stigma currently attached to AI disclosure. That reversal of incentives is the mechanism through which the flipped default would produce genuine transparency.

Proposal 3: AI-assisted peer review. JF alone permits AI to assist referees in drafting and editing review reports. Every other journal either prohibits reviewer AI use or fails to address it. The consequences are not abstract. Researchers submitting to JF are evaluated by referees who can use the same AI-assisted tools the authors used. Researchers submitting to TAR, JAE, or JFE face referees operating under an asymmetric constraint. That asymmetry does not merely disadvantage individual papers—it systematically degrades accounting-journal peer review's capacity to catch the problems AI-assisted production creates. The accounting journals' confidentiality concern is real but solvable. It is an engineering problem, not a principled objection to AI-assisted review.

A Reform Ranking: Which Journals Need to Change Most?

JF requires the least reform: it already permits unrestricted AI use, permits AI-assisted referee reports, and requires only a brief description rather than a formal declaration. Our proposals would require JF to invert the disclosure default and develop explicit copyright guidance. RFS is similarly positioned but needs explicit reviewer guidance. JFE and JAE have made meaningful progress on permitted scope but retain the reviewer prohibition and backwards disclosure default. TAR has the broadest formal scope permission of the accounting journals but the most demanding disclosure requirement. JAR requires the most reform: scope restriction, backwards disclosure, and the reviewer prohibition all need to change.

The most efficient path to field-wide reform is to engage at the publisher level. Wiley (JAR, JF), Elsevier (JAE, JFE), and OUP (RFS) each set baseline policies that their journals cannot easily deviate from without publisher approval. The AAA, as TAR's publisher and the discipline's professional association, is the one institution with both editorial and organizational standing to lead a coordinated conversation across all six journals about what the field's AI policy should look like.

4. What Should 'Intellectual Alpha' Mean Now?

The management and accounting fields are obsessed with intellectual leadership. In promotion letters, committees spend hours debating who was the "thought leader" and who was just the research assistant crunching the data. This distinction matters because tenure is, at its core, a bet

on intellectual productivity. The institution is committing thirty or more years of resources to a person. It wants to know that person can generate ideas, not just execute them.

Compare this to the hard sciences. In a physics or biology paper with twenty authors, the person who runs the lab—the one who provided the vision and the infrastructure—often shows up last on the byline. They recognize that knowledge is a collective output of a system, not the solo performance of a lead actor. AI breaks the management model entirely. When ideas are produced in loops involving multiple humans and machines, attribution becomes both difficult and conceptually misplaced. The unit of production is the loop. The idea is a product of the interaction.

As AI handles what was previously research assistant work—data cleaning, literature synthesis, first drafts—our definition of intellectual alpha must shift. Alpha should not be about being the person who sat alone in a room and typed fifty pages of unreadable prose. It should be about the ability to act as a coherence anchor: the one who orchestrates the AI agents, recognizes the signal in the noise, ensures the output is grounded in genuine domain expertise, and takes responsibility for whether the paper actually advances knowledge. We call this the Recognizer function: the ability to identify, from among everything a HIT can produce, what actually constitutes genuine knowledge.

There are two thresholds the Recognizer is working across. The first is whether the loop's output is ready for community judgment—whether the claims are traceable, the process documentable, and the argument can survive adversarial scrutiny. A claim that meets these conditions is warrantable: it has cleared the production threshold. The second threshold is the verdict the academic community renders afterward—whether the work actually advances the field, gets taken up, and changes how others think and work. That is warranted knowledge. Warrantable is what the production process can achieve; warranted is what the community confers. The Recognizer controls the first. No one inside the production process controls the second.

This creates what we call the Recognizer Problem. The loop can only surface knowledge that the Recognizer is already capable of recognizing as knowledge. If the human orchestrator cannot tell the difference between a genuine reframing and a sophisticated recombination, the loop's most distinctive outputs get passed through uncritically or discarded for reasons the Recognizer cannot articulate. The loop is bounded by the person at its center. This is not an argument against HITs. It is the central argument for taking the development of Recognizer capacity seriously—in PhD programs, in mentorship structures, and in how tenure committees evaluate intellectual contribution.

What Replaces Paper Counts?

Tenure committees are not yet evaluating for coherence anchoring. They are still largely counting papers, weighted by journal rankings calibrated to a production function that no longer exists. A scholar who can produce thirty AI-assisted papers a year can game this system trivially. And some will.

There is, however, one mechanism that already exists and has been systematically undervalued in tenure evaluation: the live academic presentation. Workshops and conference presentations are the closest thing the profession has to a viva voce—the oral examination that many universities still use to assess whether a doctoral candidate has genuinely understood what they are claiming to have learned. A researcher who has genuinely originated an idea—who understands the empirical design from the inside, who has thought through the identification strategy and its limits—will perform differently in a live seminar than a researcher who has supervised a capable AI through a process they do not fully own.

Tenure dossiers should record not just publications but live evaluative presentations—workshops, conferences, and invited seminars where the candidate fielded questions from an informed audience—together with written assessments from session chairs or designated discussants. We caveat that access to major workshops and conferences is not uniform; regional and field-specific conferences should count, and the number of presentations required should be calibrated to what is realistic across the diversity of institutions that produce good research.

Internal faculty letters also deserve far greater weight than they currently receive. Internal faculty have observed how their junior colleagues think across hundreds of seminars—not just seminars where the junior scholar presents their own work, but every seminar where they ask a question, challenge a speaker, or sit silently when they should have pushed back. An external letter writer, however distinguished, is evaluating a curated artifact. An internal colleague is evaluating the thinking process that generated it. In the AI era, when the artifact itself may be the product of a loop that the candidate orchestrated but never fully owned, the internal colleague's testimony about the quality of that orchestration becomes the most informative evidence available. We recommend that departments institutionalize internal faculty letters as a required component of the tenure dossier—not as courtesy letters but as structured evaluations of intellectual process by colleagues who have watched the candidate think in real time.

5. Does AI Actually Produce Better Science?

This paper's argument is ultimately about organizational form. But the organizational question is subordinate to a prior one: what is all of this for? The answer, stated plainly, is better science. Not more papers, not more citations, not faster review cycles or cleaner disclosure norms. The ultimate aim of the Hybrid Intelligent Team is to produce knowledge that is truer, more useful, more consequential, and more resistant to being wrong than the artisanal alternative. Every reform we propose—the disclosure flip, the AI fourth referee, the viva voce evaluation, the coherence anchoring criterion—is justified by one criterion: does it produce better science? If HIT produces more papers but not better science, we have failed.

The evidence from other fields is instructive. The most dramatic example is structural biology. For decades, the protein folding problem was considered one of biology's grand unsolved challenges. In 2021, DeepMind's AlphaFold2 solved it with accuracy equivalent to experimental methods—a

step-function advance, not incremental progress. The AlphaFold Protein Structure Database now contains structure predictions for hundreds of millions of proteins, representing decades of potential scientific progress compressed into a single system. The pattern across comparable advances in drug discovery, astronomy, and medicine is consistent: AI's contribution to better science is about expanding the tractable question space—allowing researchers to ask questions that were previously unanswerable because the data scale was too large or the interdisciplinary synthesis too costly.

The accounting and finance community should be asking what its equivalent of the protein folding problem is. One concrete example: SEC Chair Atkins has stated publicly that risk factors in the 10-K have "become a repository for too much" and that risk-averse firms "dump in the kitchen sink," producing disclosures that are not read and do not serve investors. The standard value-relevance studies—which correlate stock prices to a piece of accounting information—are routinely dismissed by practitioners and policymakers as statistical associations without causality. They do not answer the chair's challenge. Can AI and HIT help us understand, at scale and with causal credibility, which disclosures investors actually use, which risk factors actually shift capital allocation, and which are mere compliance theater? If accounting researchers cannot answer that question better than they have, the field's claim to policy relevance is difficult to defend.

6. Who Would Not Get Tenure Under the Proposed System?

The preceding argument has an uncomfortable implication. If the tenure criterion shifts from publication count to usefulness, from artisanal execution to coherence anchoring, and from field-specific stylistic conformity to genuine interdisciplinary synthesis, some profiles that have historically earned tenure at top schools would not do so under the new framework. It is worth being specific about who.

The Artisanal Technician. The most common tenure profile in empirical accounting and finance is the scholar who has mastered a particular methodology—regression discontinuity design, natural experiments using regulatory changes, textual analysis of disclosures—and applies it, carefully and competently, to question after question. The work is technically rigorous, reproducible, and publishable. It is also, increasingly, AI-replaceable in its core execution. Technical facility without ideational originality is, in the HIT era, execution without intellectual alpha.

The Journal Game Player. A second profile that would struggle is the scholar who has optimized not for truth but for acceptance—who has learned the stylistic and rhetorical conventions of target journals with sufficient precision to produce papers that clear editorial filters without necessarily advancing understanding. This is not dishonesty. It is rational adaptation to the incentive structure the current system created. But AI now exposes it. If an AI can translate any clear argument into AMR-speak in thirty seconds—as demonstrated in the appendix to this paper—the ability to produce AMR-speak is no longer a proxy for intellectual engagement.

The Dataset Moat Builder. Some tenure cases have been built, in part, on privileged access to proprietary data that other researchers cannot easily obtain. The value created was partly informational and partly positional. AI substantially erodes the positional component. Dataset moat builders whose insights were genuinely important—who found things in their data that changed the field's understanding—still add value. Those whose contribution was primarily access rather than insight face a harder evaluation.

The Political Scholar. A political scholar builds a career not by asking important questions or finding important answers, but by navigating the preferences and tribal loyalties of the editors who control the gates. AI dismantles this model in two distinct ways: first, AI-assisted papers search the literature far more comprehensively than any human working to satisfy a particular editor's known preferences, making strategic omission harder to sustain; and second, AI dramatically lowers the cost of testing whether the priors that have governed editorial selection are actually supported by evidence.

We hasten to clarify that none of these profiles is valueless or dishonest. Each represented a rational adaptation to the incentive structure created by the artisanal model. What we are arguing is that the artisanal model's incentive structure is being made obsolete, and the profiles it selected for are least likely to generate the kind of usefulness that justifies reduced teaching loads, expensive research infrastructure, and the institutional prestige of academic scholarship.

7. Has the Artisanal Meritocracy Broken Down?

The artisanal model of academic research selected for a specific personality type as surely as any structured hiring process. The scholar who survived a six-year PhD, navigated the job market, and then spent the next six years producing the three to four papers needed for tenure at a top school had to possess, above all else, a specific kind of persistence. The ability to work on a single empirical project for two or three years, absorb rejection after rejection, revise in response to contradictory referee reports without losing the thread, and return to the same dataset week after week until it yielded—this was the selection condition for academic survival.

The correlation between persistence and output quality is breaking down. In a world where AI can perform the execution functions that persistence once unlocked—the literature synthesis, the code, the robustness checks, the translation into target-journal style—persistence is no longer reliably correlated with output quality. A scholar with genuine ideational capacity and a functioning Hybrid Intelligent Team can produce, evaluate, and iterate faster than a highly persistent but less creative rival who has not learned to work with AI.

More important, the scholars who will thrive in the HIT era are not primarily those who work hardest in the artisanal sense. They are those who think best. The kind of persistence that matters has changed: the refusal to accept a shallow question when a deeper one is available; the willingness to iterate through AI-generated outputs until the finding is genuinely right rather than

merely publishable; the discipline to evaluate the loop's output against the standard of truth rather than the standard of reviewer satisfaction.

The institutional challenge is that the scholars who currently sit on tenure committees are overwhelmingly products of the artisanal selection process. Asking them to evaluate candidates by criteria that discount the virtues through which they themselves succeeded requires a degree of institutional self-awareness that does not come naturally to any profession. We are not arguing that the old virtues were fraudulent. We are arguing that they are no longer sufficient, and that rewarding them in the absence of the ideational capacity that was always the underlying goal is now an error with observable costs.

Curiosity, Entrepreneurship, and the System Cracker Problem

The persistence framing misses something important: the difference between scholars who are curious and scholars who are strategic. The artisanal system was supposed to reward the former. It has largely rewarded the latter. If AI collapses the execution cost to near zero, the curious researcher is freed to pursue the ambitious question they previously could not afford to risk. A failed three-year project in the artisanal world was a career-threatening cost. A failed three-month project in the HIT world is a sunk cost you can walk away from and try something new. The option value of curiosity has increased dramatically. Conversely, the system cracker gains less from AI: the strategic value of knowing which questions editors like and which workshops to attend has not increased.

8. Does AI Enable Genuine Interdisciplinary Synthesis?

The most significant—and least discussed—implication of AI for academic research may be what it does to interdisciplinary borrowing. Accounting's adoption of economics and psychology transformed empirical accounting research in ways that purely internal development could not have achieved. Finance's engagement with cognitive psychology produced behavioral finance. But interdisciplinary borrowing was expensive.

Studies of research funding and citation patterns have found that interdisciplinary papers face higher rejection rates, longer review times, and slower initial citation accumulation than field-specific work, even when controlling for quality—a finding that reflects the mismatch between interdisciplinary production and discipline-specific evaluation infrastructure. AI has collapsed this cost structure. A researcher who wants to understand whether mechanism design insights from contract theory might illuminate audit partner assignment can now engage with an AI system that has processed the entire relevant literature across all relevant fields. The synthesis that previously required a five-person team spanning three departments can now be initiated—not completed, but initiated—in a focused session.

The implication for tenure evaluation is direct: committees should not require that a candidate who has produced genuinely illuminating interdisciplinary synthesis also satisfy the full methodological standards of the source discipline. They should ask whether the synthesis advances understanding of an accounting or finance problem. The value of the transfer is measured at the destination, not the origin.

9. What Should Count as a Tenurable Contribution Now?

The preceding arguments converge on a question that tenure committees will soon be unable to defer: what does it mean to make a tenurable contribution to knowledge when the cost of producing research has collapsed and the proxies for intellectual contribution have been stripped away? We propose five criteria.

First, the contribution must identify a question that needed answering. This is the ideation criterion, and it is the most important and least measurable of the five. The criterion is importance: would someone who needed to understand the world be better off knowing the answer to this question? In a world where producing a competent answer is cheap, the scarcity that matters is in the question.

Second, the contribution must demonstrate genuine epistemic judgment. The coherence anchoring concept describes part of this: the ability to orchestrate the analytical process, evaluate its outputs against the standard of truth, and take responsibility for the conclusions. This capacity cannot be read off a CV. It is visible to co-authors, doctoral students, and workshop audiences who have watched a researcher reason through a hard problem in real time.

Third, the contribution must be useful to someone who needs it. Usefulness—the capacity to change how practitioners, policymakers, or other researchers understand and navigate the world—should replace publication count as the primary research criterion. The measure of usefulness is not citation count, which is increasingly gameable. The best measures are domain-specific: did the paper change how auditors, CFOs, or regulators approach a practical problem? Has it been cited in a way that indicates its results were actually used, not merely acknowledged?

Fourth, the contribution must demonstrate intellectual honesty about what AI did. A researcher who cannot defend the assumptions on which their theoretical framework rests, who cannot articulate why the question they pursued was worth pursuing, has failed the coherence anchoring test regardless of how impressive the output looks. A tenure dossier that includes an honest account of what human and machine members of the team each contributed is not a confession. It is evidence that the scholar understands their own process.

Fifth, the contribution must survive the live encounter. The viva voce reform—the formalization of workshop and conference presentations as evaluative evidence—reflects a simple and robust insight: genuine understanding of a research contribution cannot be faked in real time. The questions a good workshop audience asks are precisely the questions that separate the scholar

who has genuinely originated an idea from the one who has supervised an efficient production process without fully internalizing what it produced.

10. Is Journal Style a Genuine Quality Signal or Tribal Gatekeeping?

One of the most striking barriers in our field is the obsession with journal style. Journals in different fields have developed idiosyncratic conventions—what tense to use, how to structure an argument, how much to theorize—that function less as quality standards and more as tribal markers. Papers that do not sound like papers from that field get rejected before the quality of their ideas is ever evaluated.

We have seen field research on the PCAOB that one of us co-authored rejected at a top management journal not because the data was bad, but because it was not written in ASQ or AMR style. This is gatekeeping masquerading as rigor. These styles often produce papers that are fundamentally unreadable to the very practitioners—CFOs, auditors, regulators—who most need the insights. AI exposes this absurdity completely. If an AI can take a clear, practically useful paper and translate it into AMR-speak in thirty seconds—as demonstrated in the appendix to this paper—what is that style actually worth? The right response is to stop treating stylistic conformity as a quality signal and judge papers on their ideas, evidence, and argument.

11. Can an AI Referee Handle Interdisciplinary Papers Better Than Humans?

Publishing interdisciplinary work is currently a nightmare. The most important questions—about sustainability, financial systems, corporate governance, public health, inequality—do not sit inside any single discipline. And the academic system is extraordinarily bad at handling them.

If you send a paper covering accounting, AI, and organizational behavior to four human experts, they will often reject it because it does not fit the narrow silos of their respective sub-fields. Each reviewer evaluates their domain rigorously and the rest impressionistically. The paper falls through the cracks. AI refereeing changes this. An AI referee can assess coherence across domains—whether the accounting argument is consistent with the organizational behavior literature, whether the AI claims are grounded in what the technology actually does—rather than whether the paper satisfies the tribal rituals of one department.

Our Columbia colleague Oded Netzer proposed a practical intermediate step: include AI as a fourth referee on every paper. There would be no obligation to respond to it. Authors and editors could ignore it entirely. But it would socialize the idea of AI as an objective participant in the evaluation process, and over time the system would accumulate data on how AI assessments correlate with eventual outcomes—which is genuinely useful information for improving the process.

12. The Coming Flood: Can We Manage 80 Percent Rejection Rates?

We must be honest about what happens next. When the production of knowledge is accelerated this dramatically, we will be inundated with AI-enhanced papers. Most of them will be worthless. One of us is a senior editor at Management Science and received upward of 350 submissions in 2025. AI is making this dramatically worse. The tools that make it easy to produce AI-assisted papers are the same tools that are prohibited—by journal policy—for use by reviewers. The result is a system under asymmetric pressure: production is AI-assisted at scale, and review remains a human obligation conducted under guidelines that prohibit using the same tools. This is not a temporary adjustment problem. It is a structural consequence of asking the wrong institutional question.

We expect desk rejection rates to climb toward 80 to 90 percent at top journals—this is our projection, not a measured rate, but it is consistent with the trajectory visible in submission volumes and the ease of producing superficially competent AI-assisted work. The good news is that AI can handle those rejections. If a paper is a mere recombination of existing literature with no new data or genuine coherence anchoring, an AI can identify that immediately. This frees up human editors and referees to focus on the 10 percent of papers that actually move the needle. The constraint is no longer production. It is attention.

13. Teaching Loads and the Released-Time Question

This brings us to the most sensitive topic: teaching loads. Currently, untenured and some tenured faculty often teach 2.5 or 3 classes per year while adjuncts teach six. That 3.5-class subsidy is meant to provide time for the artisanal research process that used to take three years. The argument is that the research produced creates value exceeding the teaching foregone. But if a Hybrid Intelligent Team can produce a high-quality, warranted paper in three months—or even one—what do we do with the released time? If we continue to produce papers at the old artisanal pace while using 2026 technology, we violate the trust of the students and taxpayers who fund our research time.

We have a choice. Three options are most plausible. First, more and better teaching—not just more courses, but AI-transformed pedagogy that enables more interactive, more responsive, more personalized teaching that reaches students in ways that large lectures cannot. Second, more and better mentorship of junior researchers: the skills that matter most in the AI era—judgment about what is worth pursuing, ability to evaluate AI outputs critically, deep domain expertise that AI cannot replicate—are transmitted through apprenticeship, not through reading papers. Third, more ambitious research—questions too large or too risky to pursue under the old production function but tractable when the cost of iteration is lower. We are not arguing that teaching loads should simply be increased. We are arguing that the case for reduced loads has to be reexamined honestly.

14. Is Recognition, Not Production, the Real Bottleneck?

AI accelerates production. It does not replace judgment. And the supply of credible Recognizers—scholars with the domain knowledge, the tacit understanding, and the accountability to evaluate whether a knowledge claim is genuinely warrantable rather than merely fluent—is not scaling with the volume of output that AI-assisted production is now capable of generating. The bottleneck in knowledge production has shifted from creation to recognition. We will not run out of papers. We will run out of credible evaluation.

This argument does not conflict with our proposal for an AI fourth referee—but the distinction matters. The fourth-referee proposal operates at the level of volume management: identifying papers that are mere recombinations of existing literature, that lack original data, that fail basic internal coherence tests. This is mechanical filtering of obvious noise at scale. The Recognizer Problem is a different claim. It says that once a paper has cleared the noise filter, the judgment of whether it constitutes genuine knowledge remains irreducibly human. AI can filter out the worthless papers. It cannot yet identify what makes the worthy ones worth anything.

The tacit knowledge dimension is not abstract. The machines can generate text that is fluent, accurate on retrievable facts, and structurally coherent. What they cannot do is produce text that is right in the way a domain expert's judgment of rightness differs from a checklist of correctness criteria. That gap between accurate and right is where tacit knowledge lives, and no AI yet bridges it.

15. What Does the HIT Era Mean for PhD Programs and Junior Faculty?

The Ideation Bottleneck Was Always There: The Real Question Is the Resourcing Calculus

Identifying ideation as the thing that actually matters in research is not news. What changes in the HIT era is not the diagnosis. It is the resourcing calculus. If AI makes execution largely replicable, there is, in theory, a pool of attention and time previously committed to methods training that could be reallocated toward the deliberate cultivation of ideation capacity. Whether programs will actually make that reallocation is a different and harder question. What AI is most likely to produce, absent deliberate institutional intervention, is not better PhD students asking more important questions. It is more PhD students producing more shallow papers faster.

The Domain Depth Problem: A Pre-Existing Condition That AI May Worsen

The Recognizer function is not a general-purpose cognitive skill. It is domain knowledge applied to the evaluation of AI-generated output. An AI system can synthesize a literature, identify gaps, and generate hypotheses at scale. The judgment of whether a synthesized gap is a real gap, whether a generated hypothesis has already been answered in a literature the AI has misread, whether an identified anomaly is a genuine feature of the world or an artifact of measurement—all of this

requires the kind of internalized domain understanding that no amount of prompt engineering can substitute for.

The structural problem with accounting PhD programs is that the field has historically under-invested in genuine domain depth, prioritizing methodological sophistication instead. Students learn to run regressions before they understand what accounting is actually for. They learn identification strategies before they have developed convictions about which questions are important. If AI frees up resources previously assigned to execution, the question of what fills that space is the most consequential question the discipline faces.

The Bootstrapping Problem: Can You Be the Recognizer Without First Being the Executor?

This is the most fundamental complication and has received the least attention. We call it the bootstrapping problem. Domain expertise, at the doctoral level, has historically been acquired in part through the process of mastering a method. When a doctoral student spends two years learning econometrics, she is not merely learning to run regressions. She is learning how data is generated, what the assumptions embedded in an estimator actually require, and what distinguishes a robust finding from a fragile one. Understanding why a difference-in-differences design requires parallel trends is not a piece of knowledge that can be decoupled from the experience of testing whether parallel trends actually hold in a dataset.

If doctoral programs delegate execution-type tasks to AI before students have built that foundational understanding, we risk producing PhD graduates who can operate AI systems they cannot critically evaluate. The Recognizer requires someone who has, at some point, been the executor. This does not mean that methods training should continue unchanged. It means that the goal of methods training must be reframed: not to produce students who can run a clean regression, but to produce students who can evaluate whether an AI-generated analysis has made the right assumptions, chosen the right estimator, and interpreted the right coefficient.

The Atrophy Problem: What Happens to the Cognitive Muscles AI Is Replacing?

There is a more disturbing version of the PhD training argument. The bootstrapping problem applies not just to econometric methods but to writing and reading themselves. When a doctoral student drafts a paper—struggling to translate a half-formed empirical intuition into a structure that can survive the scrutiny of a hostile referee—something is happening that is not merely about producing text. The act of writing forces the writer to discover what they actually think. Arguments that seemed coherent in conversation collapse when they have to be committed to a paragraph. Writers do not first think and then write. They think by writing. A PhD student who delegates first-draft generation to AI before they have internalized this discipline is not saving time. They are outsourcing the cognitive process through which understanding is actually built.

The risk is not that students will produce bad papers—AI, properly prompted, produces fluent, structurally coherent prose. The risk is that students will lose the capacity to evaluate whether the argument the prose contains is actually good. The Recognizer function requires precisely this evaluative capacity. But the evaluative capacity is built through the experience of failing at the same task. You learn to spot a weak argument by having written weak arguments and been shown where they failed.

The reading problem is already visible. Doctoral students who struggle to read a forty-page paper with sustained concentration are not merely inconvenienced. They are impaired in the precise activity that domain depth requires. Genuine familiarity with a literature—the kind that lets a Recognizer immediately identify when an AI-generated synthesis has misread a key paper—cannot be acquired by reading AI summaries. It is acquired by reading the papers, in their full argumentative structure, with sufficient patience to notice what the summary omits.

What Are Human PhDs For?

We believe the answer is coherence anchoring and ideation. More concretely, moving up the value chain in the HIT era requires four things, in ascending order of difficulty.

First: learning to work in the loop without being captured by it. The minimum viable HIT skill is the ability to engage with AI output critically rather than receptively. This is a teachable skill, and it is the entry-level requirement for HIT-era research. PhD programs should train it explicitly, through exercises in which students are given AI-generated analyses and asked to find the error—not because the error is hidden, but because identifying it requires understanding the domain well enough to know what right looks like.

Second: building domain depth through reading that cannot be shortcut. The irreducible core of PhD training in the HIT era is not methods mastery and is not AI fluency. It is genuine domain knowledge—the kind that comes from reading the literature slowly, arguing about it in seminars, and developing enough internalized understanding to have real convictions about which questions matter. The measure of success is not whether the student can summarize fifty papers. It is whether the student can explain why fifteen of those papers were wrong about something.

Third: writing as a thinking discipline, not a production task. Programs should require doctoral students to write before they know what they are trying to say—to use writing as a tool for discovering the argument, not for transcribing one already formed. AI should be used after the student has produced a first draft, not before. The student's first draft, however bad, is evidence of their current understanding. This sequencing preserves the cognitive function of writing while taking advantage of AI's capacity to identify gaps and inconsistencies.

Fourth: developing the judgment to identify questions that matter. Programs should require doctoral students to regularly answer, in writing, the question "who would be better off knowing

this, and how?" for every project they consider. The discipline of answering that question honestly is the discipline of ideation.

Implications for Junior Faculty in Transition

The foregoing analysis has direct implications for junior faculty who completed their training before the AI transition and are now navigating a tenure system that has not yet decided what it is selecting for.

The first concern is the quantity trap. Under the artisanal model, the tenure clock was calibrated to the production time of four to six papers at top journals. Under the HIT model, the cost of producing papers has fallen substantially. If departments respond by simply raising the bar—expecting more papers over the same period—the efficiency gains of AI accrue entirely to the institution, while the junior scholar bears the cost of an escalating standard without a corresponding increase in time for deeper thinking.

The second concern is the evaluation mismatch. Junior faculty who are strong coherence anchors and genuine Recognizers may not look impressive under traditional evaluation metrics. Their CVs may show fewer papers than colleagues who have learned to use AI-assisted production at scale to generate competent but unimportant work. Tenure committees that are not explicitly evaluating for ideation capacity and epistemic judgment will systematically undervalue these scholars.

Three structural responses follow. First, departments should adopt explicit HIT-era evaluation criteria at the point of hiring, not at the point of promotion. Second, mentorship in the HIT era should focus on what AI cannot provide: domain depth, judgment about which questions are worth pursuing, and the live intellectual presence that the viva voce criterion is designed to assess. Third, departments have an obligation to evaluate junior scholars trained in the artisanal model with appropriate generosity for the transition period.

16. What Have We Probably Missed?

Any attempt to describe what AI is doing to knowledge production will miss things. We acknowledge three areas of genuine uncertainty.

Tacit knowledge. Much of what makes a great researcher great cannot be articulated, transmitted through prompts, or replicated by AI. The ability to recognize an important empirical anomaly, the judgment about which detail in an interview reveals the deepest truth, the intuition that a theoretical framework will not survive contact with data in a particular domain—these are capabilities built through years of experience and immersion in a field. We worry that our proposal for unrestricted AI use may inadvertently accelerate the production of work that is superficially sophisticated but lacks this tacit dimension.

The small-school problem. Access to major workshops and conference venues remains genuinely unequal across institutions and geographies. Having said that, the access challenge is largely misdirected when applied to AI tools themselves. The frontier AI tools are available for costs that are trivially low by any measure of academic resource. The subscription costs involved are on the order of what researchers spend on a single conference dinner. The barriers that matter in the HIT era are not access barriers. They are time barriers and mentorship barriers. Professional associations should focus their equity efforts there.

The replication crisis. Academic research has been contending with a replication crisis for over a decade. AI could make this better or worse: better, because AI can check the logic of an analysis and identify potential specification errors before submission; worse, because AI can make it easier to produce sophisticated-looking analyses that cannot withstand independent scrutiny. We lean toward thinking that having AI independently attempt to replicate a result before submission could substantially improve replication rates. We caveat that this is speculative.

Conclusion: Stop Obsessing Over the Byline

The production of knowledge is no longer a human-only endeavor. The Hybrid Intelligent Team is the new reality. We can either fight it with outdated disclosure rules and artisanal pretenses, or we can embrace it to solve real problems in accounting and finance. We are moving from knowledge produced by individuals to knowledge produced by systems—systems that include humans, machines, and the loops connecting them.

The institutions that organize knowledge production—journals, tenure systems, teaching load structures, disclosure norms—were not designed for this world. The question is not whether they will change. It is whether the changes will be deliberate, with forethought about what we are trying to preserve, or reactive and incoherent. The naive and reckless position is to pretend the current system is working. Submission volumes are rising. Review quality is declining. Interdisciplinary work is being suppressed by stylistic gatekeeping. Junior scholars are learning to hide their AI use rather than develop it. Tenure committees are evaluating candidates by metrics increasingly disconnected from intellectual contribution or its usefulness to policy and practice.

What we are really arguing for is honesty—about what AI does and does not do, about who is accountable for what in the production of knowledge, about what we are trying to reward when we give someone tenure, and about what journals are actually for. The theoretical core is this: a well-run Hybrid Intelligent Team can produce warrantable outputs—claims that meet the process conditions for genuine knowledge—at a scale and speed the artisanal model never could. But warrantable is not the same as warranted. The community still has to render its verdict. And the community can only render a credible verdict if it has enough skilled Recognizers—scholars with the domain depth, the tacit judgment, and the accountability to evaluate what the loop has produced.

Flip the disclosure default so AI use is visible where it matters. Allow AI-assisted review so the evaluation apparatus keeps pace with production. Evaluate tenure candidates on epistemic judgment and live intellectual performance, not paper counts. Develop Recognizer capacity in doctoral programs before the gap between production and evaluation becomes unbridgeable.

Let us stop obsessing over the byline and start focusing on the impact.

Appendix: The Style Trap

To illustrate the absurdity of modern journal requirements, we present two versions of the introduction to our recent research on the effectiveness of the PCAOB. One is written in direct, practical language; the other is translated into the theoretical idiom expected by top management theory journals.

Version A: The Original

> *Since the founding of the PCAOB in response to massive accounting scandals like Enron and WorldCom, there has been a polarizing debate about its efficacy. Advocates claim the regulator has enhanced audit quality and prevented fraud, while critics argue the agency is politicized, costly, and wasteful. Existing academic work on this topic suffers from two major limitations: the lack of publicly observable benchmarks for audit quality and the difficulty for outsiders to replicate research based on internal PCAOB data. In this paper, we overcome these limitations by conducting in-depth interviews with 24 senior executives—including audit partners, CFOs, and investors—to capture their lived experiences. Our goal is to provide a nuanced perspective on whether the PCAOB is truly "fit for purpose" in today's capital markets.*

Version B: The AMR/ASQ Translation

> *This study interrogates the recursive relationship between regulatory inspectoral ritualism and the professional identity of the auditor within the contemporary capital market ecosystem. Drawing upon a neo-institutional lens and the socio-materiality of oversight, we theorize the PCAOB not merely as a monitoring mechanism but as a panoptic site of "performative isomorphism." We argue that the "regulatory gaze" reconfigures the auditor's ontological security, potentially leading to a "defensive decoupling" of audit practice from its normative epistemic foundations. By employing a qualitative multi-case analysis of elite practitioners, we move beyond the "black box" of archival proxies to surface the latent tensions between bureaucratic compliance and the substantive mitigation of information asymmetry.*

Version B provides no more intellectual value than Version A. Both describe the same 24 interviews, the same finding about defensive auditing, the same critique of the PCAOB's effectiveness. What changed is the signaling vocabulary. Version B was produced, with modest human editing, in approximately thirty seconds using AI. A journal that rejects Version A on

stylistic grounds while Version B would pass the initial screen is making a decision about conformity, not substance. The style is not the knowledge. The knowledge is the knowledge.

Methodological Appendix: How This Essay Was Produced

A. Purpose and Scope

A conventional methodological appendix documents research design, data sources, and analytical procedures. This appendix does all of that, but it also treats the production process of the essay itself as primary data. The essay's central argument is that the Hybrid Intelligent Team is the new organizational form of knowledge production—that the loop connecting human and machine agents is the right unit of analysis, and that attribution to any individual node within the loop is increasingly both impossible and beside the point.

B. The HIT Configuration

The Hybrid Intelligent Team that produced this essay consisted of four agents. Robert G. Eccles served as human orchestrator, coherence anchor, and final recognizer—holding the essay's argumentative arc across the entire process, making all consequential decisions about what to keep and what to change. Shivaram Rajgopal served as co-orchestrator and domain expert: the intellectual raw material originated in a conversation between the two authors, with Shiva's practitioner instincts—about the absurdity of journal style requirements, the tenure system's miscalibration, the practical implications of AI-assisted research—forming the intellectual foundation on which the essay rests. Claude (Anthropic) occupied two structurally distinct roles: one Claude instance working with Shiva drafted the essay from his prompt, while a separate Claude instance served as Bob's review, synthesis, and appendix-writing agent, with no memory continuity between the two. That absence of continuity between the two Claude instances was not a design flaw to be managed. It was the condition the theory describes: the human orchestrators were the only continuous thread across the entire production process, which is precisely what the coherence anchoring function requires. Gemini Pro served as adversarial reviewer, receiving the essay via blind relay and returning a substantive adversarial review that identified five logical vulnerabilities, of which three were judged genuine and consequential.

C. What the Loop Produced

The most important test of any claimed loop-level process is whether the loop produced something that no individual node produced alone. Four substantive additions to the essay emerged directly from the review process: (i) the accountability asymmetry paragraph, acknowledging that AI has no professional stakes and arguing that this asymmetry strengthens rather than undermines the paper's position; (ii) the Recognizer Paradox resolution, distinguishing between AI's capacity for mechanical noise-filtering at scale and the human judgment of genuine epistemic contribution; (iii) the coherence anchoring operationalization, providing a preliminary sketch of what evaluating for

coherence anchoring would actually look like in practice; and (iv) the small-school problem discussion, revised to note that AI tool access is now near-universal and effectively free, shifting the equity concern from access to time and mentorship. None of these additions are editorial polish. They are conceptual advances that changed the essay's argument. The loop produced them. No single node did.

D. Attribution and the Essay's Own Argument

The byline names Bob Eccles and Shivaram Rajgopal. That is correct and it is not merely conventional. Human authors are named on bylines because human authors are accountable—they can be questioned, challenged, corrected, and held responsible for claims that turn out to be wrong. AI systems cannot. The asymmetry is real and the byline reflects it. But the honest answer to the attribution question, considered epistemically rather than institutionally, is that the essay belongs to the loop. The ideas that give it its distinctive character were not in any individual participant's prior framework. They emerged from the iterative process of generation, adversarial challenge, analysis, and revision that is exactly what the essay describes as the loop's constitutive function. The essay is its own evidence. The loop produced it. The humans are responsible for it. Those two statements are both true, and they are not in tension.