

Commentary

The Right to Be Called an Author: Generative AI, Epistemic Responsibility, and the Duty to Defend

Victor Frimpong¹

1. Management Department, SBS Swiss Business School, Switzerland

Generative AI has intensified scholarly publishing's need to distinguish formal accountability for a manuscript from the epistemic capacity to answer for it. Existing authorship guidance assigns responsibility to human authors regardless of AI involvement, but assigning responsibility does not establish that it can be meaningfully exercised. This paper addresses that gap by distinguishing nominal from substantive authorship responsibility and introducing the Duty to Defend (DtD): the obligation, internal to the status of scholarly authorship, to retain sufficient epistemic command of one's work to provide a reasonable account of its central claims, methods, evidence, and interpretations when legitimately challenged. The paper derives this obligation from the reciprocal structure of scholarly credit rather than treating it as an additional compliance requirement and distinguishes it from an AI-detection standard: the relevant question is never whether AI was used, but whether the named author remains capable of standing behind the contribution attributed to them. A conceptual model then identifies three configurations — integrated, attenuated, and displaced authorship — describing how technological assistance can preserve or weaken the connection between credit and human intellectual agency. The paper translates this framework into a proportionate, tiered approach to editorial assurance and examines its own limits, including evidentiary difficulty, unequal capacity to demonstrate understanding, and the risk of converting a safeguard for authorship into a mechanism of suspicion.

Correspondence: papers@team.qeios.com — Qeios will forward to the authors

1. Introduction

Scholarly authorship has traditionally connected intellectual credit with responsibility and accountability for the resulting contribution^{[1][2]}; those whose names appear on academic work possess sufficient intellectual command of that work to stand behind it. Authorship confers credit, but it also entails responsibility for the integrity and accuracy of the scholarship presented^[2]. Generative artificial intelligence (AI) makes visible a question that conventional authorship arrangements have largely left implicit: whether accepting responsibility for scholarly work necessarily means possessing sufficient epistemic command to exercise that responsibility meaningfully.

Generative AI complicates this relationship as large language models assist with language, presentation, literature synthesis, argument development, interpretation, and scholarly prose. The key issue shifts from whether AI contributed to a manuscript to whether the human author maintains intellectual control over reasoning and claims. Recent scholarship recognises this distinction: Koo and Lu^[3] state responsible authorship requires defending the reasoning behind claims, while Nascimento and da Silva^[4] stress that disclosure and human review are insufficient for responsible AI-assisted authorship. Accounts now view authorship under AI in terms of relational responsibility, epistemic stewardship, and human answerability rather than textual output alone.

Publication policies have responded primarily through disclosure, transparency, and human accountability. The International Committee of Medical Journal Editors (ICMJE)^[2], the Committee on Publication Ethics^[5], and the World Association of Medical Editors^[6] assert that AI systems cannot qualify as authors owing to their incapability to assume responsibility for scholarly work; therefore, accountability for AI-assisted content remains with human authors. These principles delineate a crucial boundary; however, assigning responsibility does not inherently establish the capacity to exercise it. An author may approve a manuscript, disclose AI utilisation, and formally accept accountability, yet lack sufficient understanding of its core theoretical reasoning, methodological choices, evidential bases, or interpretations necessary for elucidation.

This possibility suggests a distinction between nominal authorship responsibility, where responsibility is formally acknowledged, and substantive authorship responsibility, where the author has adequate epistemic authority to exercise that responsibility when the scholarly work is legitimately challenged. Research in this area is advancing. Koo and Lu^[3] introduce a Reasoning–Process–Traceability assessment focused partly on an author's capacity to justify their reasoning, whereas Van Zoonen et al.^[7] redirect

focus from AI disclosure towards accountability for claims and the capacity of human researchers to reconstruct and defend scholarly assertions. Less clear is the normative question of why defensibility itself should constitute an obligation arising from scholarly authorship.

This paper presents the Duty to Defend (DtD), a normative obligation for authors to adequately explain their main claims, choices, evidence, limitations, and contributions when challenged. DtD does not require writing every sentence alone, memorising references, or having identical expertise in collaboration. It is not an AI-detection tool or routine justification. Instead, scholarly responsibility matters only when authors can sufficiently account for their contributions.

Building on this foundation, we propose the Defensibility Principle: scholarly authorship's legitimacy is not based solely on an author's exclusive creation of the text. Instead, it depends on their ability to assume epistemic responsibility and offer a reasonable scholarly defence of the substantive contribution attributed to them. Therefore, defensibility is distinct from contribution, disclosure, accountability, and traceability, which pertain to participation, transparency, responsibility, and reconstructability, respectively. Defensibility specifically questions whether the individual claiming authorship has enough epistemic authority to provide a credible scholarly account of their claimed contribution.

Generative AI heightens the importance of this issue because it can undermine the traditional link between textual skill and evidence of intellectual authority. Heavy AI support can still align with strong epistemic authorship, while minimal or no AI use does not necessarily mean a lack of intellectual command. Here, AI is not a sign of illegitimate authorship but a stressor that reveals an existing gap between being credited as responsible for scholarship and having the epistemic ability to account for it. AI acts as the stressor; authorship remains the core focus.

The paper makes three contributions: it distinguishes nominal from substantive authorship responsibility; conceptualises the Duty to Defend as the bridge between responsibility and epistemic capacity; and advances the Defensibility Principle, exploring its boundaries for authors, editors, reviewers, and publishers. It shifts focus from AI use to what scholars must understand, account for, and defend to legitimately stand behind their work.

2. Scholarly Authorship in the Age of Generative AI

Scholarly authorship serves multiple functions, including allocating credit, identifying contributions, and assigning responsibility, but these are not always clearly linked. As collaboration and specialisation

increase, traditional authorship struggles to accurately represent contributions and responsibilities^[1]. Frameworks like CRediT address this by differentiating roles rather than treating authorship as a single category^[8].

The possibility that formal authorship attribution differs from actual intellectual contribution predates generative AI. Instances of ghost and honorary authorship illustrate that individuals may be omitted despite significant involvement or granted authorship without meeting standard criteria^[9]. Generative AI does not create the possibility of a divergence between authorship attribution and intellectual contribution; rather, it introduces a novel mechanism through which such a separation may occur.

What distinguishes generative AI is the range of scholarly activities it can mediate. AI systems assist with language refinement and organisation, but they also help synthesise the literature, develop concepts, interpret, construct arguments, and articulate theoretical relationships. The presence of sophisticated scholarly prose no longer necessarily indicates an equivalent degree of human intellectual participation in producing the reasoning expressed through that prose. This creates the possibility of a decoupling of scholarly production from epistemic command:

the production of an apparently coherent intellectual contribution without a corresponding level of understanding by the person to whom that contribution is attributed.

The distributed nature of scholarly production does not, in itself, undermine authorship. Scholarship has long relied on social, technological, and epistemic infrastructures, making exclusive individual production an inadequate basis for defining legitimate authorship^{[10][11]}. Recent accounts of generative AI likewise emphasise relational responsibility and human answerability within increasingly distributed systems of knowledge production^[10]. The relevant distinction is not between assisted and unassisted scholarship, but between forms of assistance that preserve human intellectual agency and those that may substitute for it. This distinction is crucial because AI involvement does not reliably indicate authorship quality. Extensive AI use may still require human review, and limited AI use does not guarantee that authors understand all aspects attributed to them. The main concern is not the technology itself but how it affects the author's ongoing intellectual connection to the work.

Existing publication standards primarily address this relationship through the lens of human responsibility. The ICMJE^[2], COPE Council^[5], and WAME guidelines^[6] explicitly reject AI systems as authors, citing their inability to assume the responsibilities inherent to authorship and affirming that human authors retain accountability for content generated with AI assistance. These standards uphold a

fundamental principle: technological aid does not absolve the named human author from scholarly responsibility. Nevertheless, this raises an additional inquiry.

If responsibility remains human-centric, while increasingly significant components of scholarly output are mediated through technology, what extent of human intellectual engagement is necessary for that responsibility to retain its significance?

Koo and Lu^[3] highlight an author's ongoing ability to explain the reasoning behind substantive claims. Nascimento and da Silva^[4] differentiate formal authorship responsibility from the epistemic duties involved in endorsing AI-supported claims. Van Zoonen et al.^[7] shift focus from mere disclosure to accountability for claims in scholarly work. These developments collectively indicate that while transparency about AI use is important, it alone cannot define the intellectual connection between an author and their scholarship^{[10][7]}.

This separates formal attribution, which recognises the publicly acknowledged author and associated responsibilities, from epistemic authorship, which concerns whether that individual's intellectual contribution aligns with their understanding and accountability. While they are usually aligned, generative AI highlights their potential separation. The ensuing challenge is consequently not to establish the quantity of scholarship to be generated without technological aid. Such a benchmark would be both historically dubious and progressively unfeasible. The more essential task is to discern what must retain its fundamentally human nature as scholarly production becomes technologically disseminated.

This paper contends that one aspect cannot be wholly delegated: the ability to respond intellectually for the substantive contribution for which authorship is asserted. Clarifying why this capacity derives from scholarly responsibility necessitates transitioning from formal accountability to epistemic responsibility.

3. From Formal Accountability to Epistemic Responsibility

An author might accept responsibility for a manuscript despite limited understanding of how arguments were developed, evidence was chosen, or conclusions were reached — a distinction that matters when generative AI aids not just expression but synthesis, interpretation, and reasoning. Responsibility may stay human even where some intellectual tasks are delegated.

We use epistemic responsibility to describe authors' obligation to exercise due intellectual judgement over the knowledge claims they present, endorse, and incorporate into the scholarly record^[12]. This

extends beyond liability for post-publication errors to the epistemic relationship between the author and the claims for which they assume credit and accountability. Nascimento and da Silva^[4] draw a significant distinction between authorship responsibility and epistemic responsibility in AI-assisted scientific authorship, arguing that disclosure or nominal human review alone does not satisfy the requisite epistemic duties. Contemporary discussions of AI-mediated authorship similarly hold that human responsibility depends on the ongoing ability to provide justificatory answerability, not merely on occupying the formal role of author.

Epistemic responsibility should not be equated with complete individual mastery. Modern scholarship often involves specialised epistemic collaboration, where participants share knowledge, expertise, and cognitive effort. Palermos^[13] shows that epistemic responsibility in collaborative knowledge creation cannot be attributed to a single person; it must be understood through the collaboration's structure and each participant's role. A co-author responsible for statistical analysis, for instance, need not share the theoretical knowledge of someone handling conceptual development, nor should the latter replicate every technical step taken by the former.

This appeal to distributed cognition prompts an immediate objection that warrants direct confrontation.

If epistemic responsibility can be shared among human collaborators—where a contributor does not need mastery of statistical methods – why is AI held to a different standard? Why is delegating reasoning to an expert co-author unproblematic but raising concerns when AI is delegated similar reasoning?

The answer does not reside in any inherent difference between human and artificial reasoning per se, but rather in the manner in which distributed cognition among human collaborators is actually distributed: not the ability to defend a contribution, but the identity of the individual who holds that ability. When a conceptual contributor delegates statistical reasoning to a co-author, responsibility for that reasoning is not abolished; rather, it is transferred to a second duly designated author who, if legitimately challenged, can present their own account of the methodological choices assigned to them. The requirement that someone be capable of defending the contribution is fulfilled at the author-team level, because every element of the work remains within the epistemic authority of at least one accountable individual. According to Palermos's own framework, distributed cognition characterises cognitive systems whose components remain answerable participants within the system, rather than systems in which capacity simply ceases to exist.

Generative AI cannot fulfil this role not because of output quality but because it cannot bear responsibility. That position is settled^{[5][2]}. When a human delegates reasoning, defensibility is maintained because responsibility shifts to an accountable individual. When delegated to AI without critical review, defensibility is lost since no accountable agent remains. The manuscript may credit a human author, but no one holds epistemic responsibility for that reasoning.

In this context, DtD treats AI assistance no differently from human collaboration and applies the same standard consistently. The standard requires that an accountable individual has sufficient understanding of every substantive element they take credit for. Human collaborators usually meet this requirement automatically, as they can defend their reasoning. AI assistance, however, only meets this criterion conditionally, since the system providing the reasoning cannot defend it^{[14][15]}. This makes the human author's epistemic authority the key factor in AI-assisted work, unlike in purely human collaboration.

The key requirement is sufficient, not complete, epistemic control. Problems occur when delegation creates a gap between responsibility and the epistemic ability needed to fulfil it effectively. This can be understood through two types of authorship responsibility: nominal authorship responsibility, where an individual is formally recognised as an author and accepts the related accountability, and substantive authorship responsibility, where this formal role is backed by adequate epistemic capacity to exercise it effectively.

The relationship is represented conceptually as:

(1) Formal accountability + sufficient epistemic command → substantive authorship responsibility

The formulation is not a quantitative equation but captures a conceptual dependency. Formal accountability shows who is responsible; epistemic command determines if that responsibility can be an intellectual obligation rather than just an institutional status. Without the former, responsibility lacks a clear bearer; without the latter, it risks becoming nominal.

This characteristic also helps define what disclosure can achieve. It can show that an AI system was used, clarify the purposes behind its use, and enhance transparency of the production process. These are valuable functions. However, disclosure alone cannot demonstrate if an author critically assessed an AI-generated interpretation, fully understood an argument before using it, verified the evidence supporting a claim, or explained why a particular conclusion was accepted^[16]. Transparency concerns the visibility

of assistance; epistemic responsibility concerns the quality of human intellectual engagement with the output produced.

The same reasoning extends beyond generative AI. Practices such as ghostwriting, improper delegation, highly fragmented collaboration, or excessive dependence on intermediaries can similarly lead to situations where formal authorship and substantive intellectual control become misaligned. Generative AI is distinctive not because it introduces this potential, but because it enables sophisticated intellectual delegation to occur swiftly, cost-effectively, and in ways that are difficult to discern from the final manuscript. Thus, it distinctly reveals a vulnerability that traditional authorship arrangements have often implicitly tolerated.

If scholarly authorship involves genuine responsibility rather than just nominal responsibility, a further question arises. Merely stating that an author holds epistemic responsibility is not enough; it is necessary to clarify what that responsibility entails when the scholarship is challenged. A responsibility that never obliges the author to explain the claims, choices, or reasoning for which they receive credit can remain vague. The Duty to Defend is presented as the normative link between holding epistemic responsibility and being reasonably accountable for its exercise.

4. The Duty to Defend

Scholarly authorship involves significant epistemic responsibility, which should shape what authors can reasonably be held accountable for. Current publication standards affirm that human authors are responsible for work created with AI support^{[5][2][6]}. Recent studies emphasise that responsible AI-assisted authorship goes beyond formal accountability, requiring authors to be answerable, able to defend their claims and reasoning, and to maintain epistemic responsibility^{[3][4][10][7]}.

Duty to Defend is defined as:

the normative obligation from scholarly authorship to retain enough epistemic command of one's work to explain its main claims, choices, evidence, limitations, and contribution when challenged.

4.1. The Normative Derivation of DtD

The present argument does not claim defensibility itself as a previously unidentified requirement of responsible scholarship — that observation is already well established^{[3][4][10][7]}. Its contribution is to supply the normative groundwork these accounts assume but do not provide: why does the capacity for

reasonable defence follow from claiming scholarly credit, rather than being one plausible option among several available to a responsible author?

The derivation rests on three premises. First, naming an individual as author is not a neutral description of causal involvement but a public claim to scholarly credit that the work reflects that person's intellectual judgement, made on their behalf to readers, editors, and the scholarly record^[8]. Second, existing authorship norms already treat credit and responsibility as reciprocal rather than separable: an author cannot accept credit for a contribution while disclaiming responsibility for it, which is precisely why accountability is a formal condition of authorship in current guidance^{[5][2]}. Third, responsibility that can be discharged only by an act one is presently incapable of performing is responsibility in name only — a named author who cannot, if asked, account for a central claim has not been relieved of responsibility for it but prevented from exercising it.

From these premises, the obligation to retain sufficient epistemic authority follows directly: if authorship is a claim to credit, that claim is reciprocally bound to responsibility, and responsibility requires the capacity to exercise it, then retaining that capacity is not an additional burden but a condition already implicit in accepting credit.

Two alternatives might seem to avoid this obligation, and both are worth addressing directly. The first holds that an author could discharge their responsibility by disclosing the limits of their understanding — admitting, say, that they do not fully grasp why a method was chosen — rather than retaining that understanding. This does not avoid DtD; it restates it. Honest disclosure of incapacity is compatible with responsible authorship only where it is accompanied by relinquishing the corresponding claim to credit: an author who cannot justify a contribution has reason to withdraw it, not retain credit for it under a disclaimer^{[7][17]}. Disclosure without withdrawal does not dissolve the obligation to retain command — it merely admits the obligation has not been met.

The second alternative holds that the obligation is to acquire adequate command at submission, with no requirement that it persist afterwards. This is difficult to sustain: defensibility is characteristically challenged well after publication, through commentary, failed replications, and integrity inquiries^[9], so an obligation satisfied only at submission would not secure answerability when it is actually needed. Existing guidance already speaks of authors who "remain accountable," not merely those who were accountable at submission^{[5][6]}. DtD makes explicit an ongoing obligation that current norms already presuppose.

Both alternatives collapse into, rather than replace, the position defended here. The obligation to retain sufficient epistemic authority is not one normative option among others; it is what responsible authorship already requires once credit and responsibility are recognised as reciprocal.

4.2. The Defensibility Principle

The normative logic of DtD can be expressed through the Defensibility Principle:

The legitimacy of scholarly authorship depends not on an author's exclusive production of the text, but on the capacity to assume epistemic responsibility for, and provide reasonable scholarly defence of, the substantive intellectual contribution attributed to that author.

This principle distinguishes authorship from exclusive textual creation. It is essential because scholarly endeavours have historically involved distributed contributions and technological facilitation, while generative AI extends that facilitation into activities traditionally linked more directly to intellectual output^[10]. The pertinent boundary is therefore not whether assistance was involved, but whether there remains adequate human intellectual agency behind the contribution for which authorship is asserted.

DtD also differs from neighbouring concepts. Contribution pertains to what an individual has added to the research. Disclosure concerns transparency about how the work was produced. Accountability designates those who accept responsibility for it. Traceability pertains to whether relevant processes, sources, or decisions can be reconstructed. Defensibility concerns whether the person claiming authorship possesses sufficient epistemic authority to provide a credible scholarly account of the substantive contribution attributed to them. While these concepts may reinforce one another, they are not interchangeable.

4.3. What Must an Author Be Able to Defend?

The content of DtD depends on the nature of the claim. It does not require defending every sentence or fact equally. Defensibility relates to the intellectual elements that turn the contribution into scholarly knowledge. Four dimensions are especially important.

1. Conceptual defensibility evaluates whether an author can clearly explain the main argument, theoretical relationships, key concepts, and their contribution.
2. Methodological defensibility relates to justifying critical choices in research design, data, analytical methods, or other procedures within the author's scope.

3. Evidential defensibility involves understanding why key evidence or sources support the claims and recognising their limitations.
4. Interpretive defensibility assesses the ability to explain how conclusions are reached and why certain interpretations are accepted, rejected, or limited.

These dimensions are not intended as a universal checklist or measurement tool. Instead, they aim to clarify the kinds of epistemic connections that can underpin meaningful scholarly responsibility. Their significance will differ by context, whether theoretical, empirical, methodological, or collaborative.

4.4. Defensibility Is Not the Same as Correctness

An important distinction exists between defending a scholarly claim and being correct. Scholarship is inherently contestable; well-supported arguments can be rejected, findings may not replicate, and interpretations can be replaced by better explanations^{[18][19]}. DtD cannot require authors to guarantee the truth of all they publish.

Defensibility primarily pertains to whether an author can furnish a credible scholarly account explaining the rationale behind adopting a particular claim, method, interpretation, or conclusion, given the existing knowledge and evidence. A claim can be defensible yet later proven incorrect. Conversely, a claim may independently be correct, even if the individual presenting it lacks a comprehensive understanding of its correctness. This latter circumstance is especially significant in the context of authorship, as epistemic responsibility cannot depend solely on an output's incidental accuracy. DtD concerns the relationship between judgement and responsibility, rather than retrospective infallibility.

4.5. A Proportionate Duty

The Duty to Defend should align with each author's claimed contribution. Modern research often involves diverse roles, and contributorship models acknowledge that authors may contribute in various ways such as conceptualisation, methodology, analysis, data curation, writing, supervision, or other activities^[8]. It would be unreasonable to expect all authors to have the same level of expertise across all aspects of a complex collaborative project.

DtD is inherently contribution-sensitive. Authors must demonstrate sufficient mastery of the substantive contributions they are attributed with, while also comprehending the core purpose, claims, and implications of the collaborative work they endorse. A methodological specialist is not required to replicate the theoretical expertise of a co-author but should be capable of explaining the methodological

contribution for which they receive credit. Likewise, a conceptual contributor is not obliged to conduct every technical analysis but must be able to articulate the central argument associated with their name, especially when claiming substantive responsibility. Proportionality safeguards against DtD degenerating into a demand for universal expertise, and it also aligns epistemic responsibility with the distributed nature of contemporary scholarship.

4.6. DtD Is Not an AI-Detection Test

The Duty to Defend is different from inferring AI use. An author's difficulty in answering does not confirm AI use, just as fluent defence does not prove AI absence. Factors like language skills, anxiety, conventions, memory, style, disability, time, and expertise distribution influence scholarly expression.

The key question, therefore, is not:

Did this person use AI?

but:

Does this person retain sufficient epistemic command to stand behind the substantive contribution for which authorship is claimed?

This distinction is important. Extensive AI support can still align with strong epistemic authorship when outputs are carefully scrutinised, checked, revised, and integrated using human judgment. On the other hand, weak epistemic control can happen without generative AI, through ghostwriting, misdelegation, or poorly managed collaboration. DtD's core logic is technology-neutral, although generative AI makes the issue clearer.

4.7. Legitimate Challenge and the Risk of Epistemic Policing

Recognising a Duty to Defend does not grant editors, reviewers, institutions, or readers unlimited rights to question authors. A normative obligation to respond differs from arbitrary demands for knowledge on demand. Otherwise, a principle meant to uphold authorship integrity might turn into a tool for epistemic surveillance.

The legitimately challenged qualifier is essential to DtD. A challenge must stem from a genuine scholarly concern, such as major reasoning inconsistencies, failure to clarify key methodological decisions, serious doubts about sources or evidence, or contradictions between claimed contributions and actual

involvement. Suspecting AI generation based on writing style, institutional ties, geography, language, or intuition is not enough to warrant heightened scrutiny.

Where clarification is warranted, the process should also be proportionate. Authors should be aware of the substantive concern, have a reasonable opportunity to respond, and, where appropriate, be able to consult the records, data, notes, code, sources, or collaborators normally involved in scholarly work. DtD concerns epistemic command, not performance under artificial examination conditions. The characteristic is consequential: the right to be called an author cannot depend on passing an interrogation. The purpose of DtD is to specify an obligation attached to authorship, not to establish a general licence for epistemic policing.

4.8. The Non-Delegable Core of Authorship

DtD implies a deeper point: scholarly production may be extensively distributed, but not every dimension of authorship is delegable. Researchers may draw on AI systems, statistical software, editors, translators, research assistants, and other intellectual infrastructure. What cannot be transferred while substantive authorship is retained is the human author's epistemic responsibility for the contribution claimed.

This does not require personally performing every intermediate operation. It requires that externally generated material become sufficiently incorporated into human scholarly judgement that the author can meaningfully stand behind it. An AI-generated interpretation that is critically examined, verified, and knowingly adopted becomes part of responsible human scholarship. One accepted merely because the prose is convincing does not, even where the resulting text happens to be accurate. The non-delegable core is not production, but answerable intellectual agency, consistent with accounts that locate authorship in continuing responsibility rather than exclusive origination^[10], for which DtD specifies the corresponding obligation.

The proposition can be stated simply: an author need not produce every element of a work but must remain capable of standing behind the substantive contribution for which authorship is claimed. In section 5, we develop a conceptual model of how technological mediation can preserve or weaken that relationship.

5. A Conceptual Model of Epistemic Authorship

The model views AI use as not inherently harmful to authorship, emphasising the links among mediation, epistemic authority, and accountability. This aligns with theories of knowledge as distributed

among multiple agents and resources^[13], and frameworks of AI-mediated authorship that situate human authorship within distributed socio-technical systems while maintaining human stewardship^[10]. The concepts of epistemic distance and the three epistemic authorship configurations in section 5.4 build on this by showing how different forms of assistance affect the author's intellectual relationship to the scholarship.

5.1. From AI Mediation to Epistemic Distance

The presence or absence of generative AI is an inadequate basis for determining whether substantive authorship has been preserved. What matters is how technological assistance affects the relationship between the author and the content attributed to them^{[7][12]}. AI may perform extensive functions while remaining subject to active human evaluation, just as limited assistance may lead an author to adopt arguments without adequate scrutiny.

Evidence from outside scholarly publishing supports this distinction. Empirical research on AI use has associated greater reliance on AI tools with cognitive offloading and lower critical-thinking performance, suggesting that technological assistance can, under some conditions, reduce rather than merely supplement active human reasoning^[20]. Such findings do not directly establish epistemic displacement in scholarly authorship, but they provide empirical support for the possibility that intensive AI reliance may weaken the cognitive engagement on which epistemic command depends. Formosa et al.^[21] found that perceptions of a human writer's authorship, creatorship, and responsibility varied with the degree of assistance received, but not with whether that assistance came from a human or an AI system. Though the study concerns creative rather than scholarly writing, it reinforces the broader point: the theoretically significant issue is not whether AI participates, but how assistance affects the relationship between the author and the resulting work.

This paper conceptualises that relationship as epistemic distance: the separation between the intellectual content attributed to an author and that author's understanding of how it was developed, justified, and incorporated. Epistemic distance is not a measure of how much text was externally generated, but of the author's intellectual relationship to it.

AI-mediated production may increase epistemic distance but need not. An author who uses AI to generate alternative explanations, identify objections, or suggest interpretations may still interrogate, verify, revise, or reject those outputs through sustained judgement, in which case mediation can be extensive

while epistemic distance stays limited. By contrast, a small amount of generated content can produce substantial epistemic distance when central claims are incorporated without adequate understanding.

The relationship can be stated at length: AI mediation creates conditions under which epistemic distance may increase, which may in turn weaken epistemic command, though neither step is automatic.

5.2. From Epistemic Command to Substantive Responsibility

As outlined in (1) within section 3, formal accountability acquires substantive responsibility only when adequate epistemic authority is maintained. The current model situates epistemic authority between the conditions of scholarly production and the ultimate responsibility claimed by the author. When human intellectual engagement sustains sufficient epistemic authority, responsibility remains exercisable. Conversely, when such engagement is substantially diminished, a tension arises between the credit and accountability formally assigned to the author and the intellectual capacity available to support these assignments. This relationship reveals why neither disclosure nor formal acceptance of accountability, in isolation, suffices. An author may accurately disclose extensive AI utilisation and accept full responsibility yet still possess inadequate command of the substantive elements of the resulting work.

5.3. The Role of the Duty to Defend

The Duty to Defend is a normative obligation tied to retained epistemic command, linking understanding with authorship. Figure 1 shows how AI-mediated scholarly production can preserve or weaken epistemic authorship based on human engagement and epistemic control.

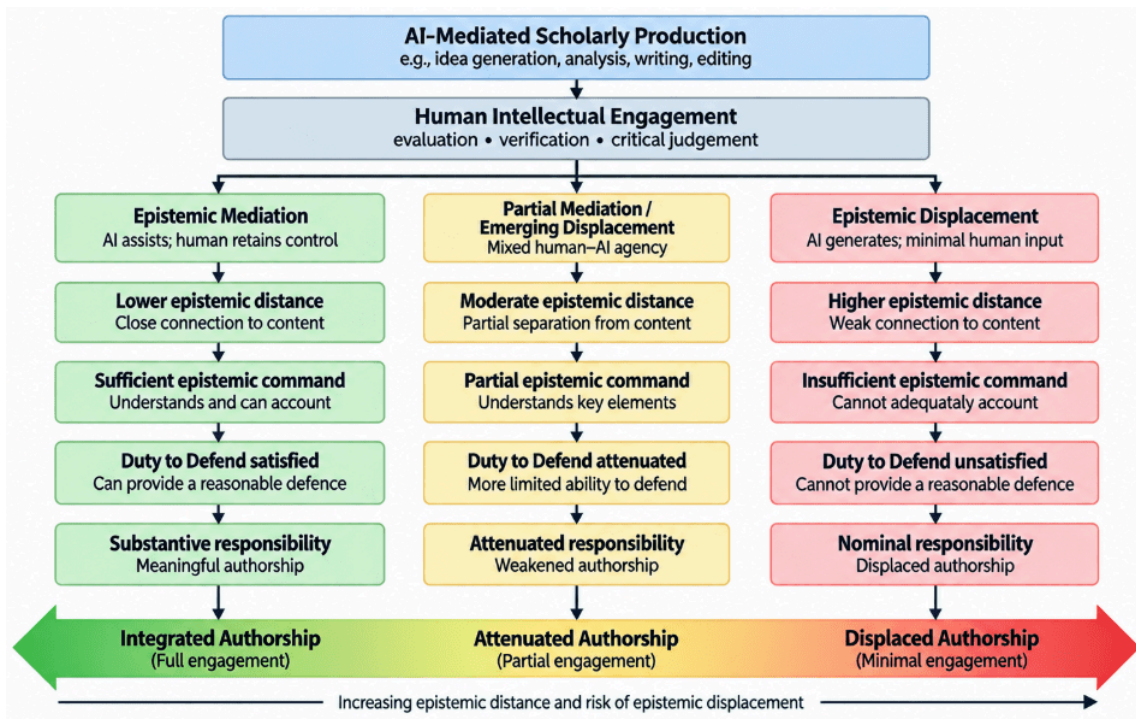


Figure 1. Conceptual Model of Epistemic Authorship and the Duty to Defend. Source: Author's conceptualisation

As shown in Figure 1, AI mediation itself does not decide who is truly an author. Instead, its importance lies in whether human involvement maintains enough epistemic authority to uphold meaningful responsibility. The various configurations are discussed below.

5.4. Three Configurations of Epistemic Authorship

External assistance relates to human intellectual agency in two broad ways: it leaves that agency intact or substitutes for it to a degree that undermines the author's capacity to account for the result. This is not a simple technological binary; the same AI function may leave agency intact in one case and substitute for it in another, depending on how the author engages with its output. But binary classification obscures the intermediate cases most common in practice, where an author retains genuine command of a contribution overall while relying substantially on externally generated reasoning in specific areas. The model distinguishes three configurations. Integrated authorship corresponds to substantive authorship responsibility as defined in section 3: external assistance is fully incorporated into the author's intellectual judgement. Displaced authorship corresponds to responsibility that is substantively nominal:

credit and formal accountability persist, but the epistemic command needed to exercise them does not. Attenuated authorship names the space between command sufficient to defend the central contribution, but with meaningful areas of unresolved epistemic distance which the ordinary/targeted/enhanced assurance levels in Table 1 (below) are designed to address.

Integrated authorship exists when technological or collaborative assistance is accompanied by sustained human evaluation and sufficient epistemic command: external contributions are incorporated into the author's judgement, so responsibility remains substantive.

Attenuated authorship exists where the author retains meaningful command of the work but shows important areas of epistemic distance, understanding the central contribution while relying substantially on externally generated reasoning, synthesis, or interpretation in particular areas. Responsibility remains possible, but the relationship between credit and intellectual agency is less secure.

Displaced authorship occurs when content attributed to an author becomes sufficiently separated from the author's understanding, rendering responsibility primarily nominal: the individual may endorse the manuscript and accept formal accountability while lacking adequate epistemic authority to explain significant components of the claimed contribution.

These are conceptual, not diagnostic, categories. They should not be translated into numerical thresholds nor used to classify authors based on isolated performances. Their purpose is to demonstrate that authorship under technologically mediated conditions varies according to the relationship between external assistance and retained human intellectual agency.

5.5. A Threshold Rather Than Mastery

The framework requires a threshold concept, not full mastery. Scholarship is collaborative, memory imperfect, expertise varies, and research often exceeds individual knowledge. The threshold is sufficient epistemic command, not exhaustive knowledge. What is adequate depends on the contribution, discipline, complexity, and responsibilities^[22]. It should stay context-sensitive and not be fixed mechanically. However, it must not allow scholarly credit to be separated from the intellectual capacity needed to exercise that credit.

This establishes the central relationship of the model:

Contribution establishes a basis for credit. Accountability establishes responsibility. Epistemic command makes that responsibility exercisable. The Duty to Defend makes that command

reasonably answerable.

Within this framework, epistemic authorship refers to a state where academic credit and responsibility are closely linked to human intellectual effort. Generative AI does not decide whether this connection exists; instead, it changes the conditions needed to uphold it. The main theoretical issue is not whether AI can be part of authorship, but how much of the epistemic connection to humans is needed for authorship to retain genuine significance.

6. Implications for Scholarly Publishing

The Duty to Defend has significant implications for scholarly publishing, shifting the focus from AI involvement in manuscript production to whether human authors have adequate intellectual agency to assume substantive responsibility for the resulting scholarship. This shift does not require replacing existing mechanisms such as peer review, disclosure, contributorship, or research integrity. Instead, it advocates interpreting these mechanisms in a way that maintains the link between scholarly credit and epistemic responsibility.

6.1. Implications for Authors

For authors, the central implication is that responsible AI use cannot be reduced to compliance with disclosure requirements. Existing guidance already places responsibility for AI-assisted content on human authors and requires them to verify generated material, ensure appropriate evidential support, and remain accountable for the resulting work^{[5][6]}. DtD extends this logic by locating responsibility in the author's continuing intellectual relationship with the contribution itself.

The relevant question is therefore not how much AI an author used, but what intellectual responsibility remained human throughout. Authors may use generative systems extensively while retaining substantive authorship where outputs are critically evaluated, sources checked, reasoning understood, and final claims reflect human scholarly judgement. Disclosure compliance cannot compensate for insufficient command of the substantive arguments or interpretations attributed to the author.

This development is expected to become increasingly significant as artificial intelligence advances from serving primarily as a writing and analytical assistant to assuming a more autonomous role in research participation. Resnik, Hosseini, and Hauswald^[23], along with Resnik et al.^[24], contend that scientists employing AI research agents bear continued responsibility for supervising these systems, accurately

disclosing their contributions, and maintaining accountability for the outcomes of the research, including addressing inquiries raised by the scholarly community. This position reinforces a fundamental premise of the current framework: enhanced technological autonomy does not absolve human authorship responsibilities; rather, it emphasises the necessity of clarifying the responsibilities that must remain under human oversight.

Contributorship declarations can support this by meaningfully identifying who was responsible for conceptualisation, methodology, analysis, interpretation, writing, and other substantive activities^[8], not merely documenting participation but clarifying the domains within which particular authors claim scholarly credit and the corresponding epistemic responsibility.

6.2. Implications for Reviewers

For reviewers, DtD does not mean they must decide if manuscripts used AI. Peer review should focus on the work's quality, coherence, evidence, methodology, originality, and significance, not on speculations about its technological origin. Reviewer exchanges may nonetheless serve as incidental evidence of the relationship between authors and their work. Requests to clarify theoretical reasoning, justify methodological decisions, reconcile inconsistent interpretations, or explain evidential choices are standard components of the peer review process. When authors respond substantively, such exchanges can exemplify ongoing intellectual engagement without the need for a distinct authorship verification.

A clear distinction must be made between manuscript clarification and authorship verification. The former involves routine scholarly efforts to improve or assess the work, while the latter evaluates if the person claiming authorship has an appropriate intellectual connection to it. Authorship verification should only be pursued when significant concerns are present. DtD should not turn reviewers into investigators of authorship legitimacy.

6.3. Implications for Editors

Editors assume a more challenging role as they are tasked not only with assessing manuscripts but also with safeguarding the integrity of the scholarly record. Generative AI has the potential to create scenarios in which a manuscript appears technically refined, yet substantial inconsistencies emerge, casting doubt on sources, logic, methodology, or the author's declared contribution.

A suitable editorial approach should focus on addressing substantive inconsistencies rather than trying to detect AI use through writing style. When clarification is needed, editors could ask for an explanation

of key methodological choices, supporting documents for disputed sources, details on how an interpretation was made, or additional information about contributor roles. These requests align more with maintaining scholarly integrity than asking the author to demonstrate that specific sentences were written without AI.

The principle of proportionality remains important. Whenever possible, choose the least intrusive method to address the main issue. Authors should reserve more detailed clarifications for special cases rather than everyday practice, and they should always be backed by clear reasons. Authors should also have a chance to respond to these clarifications to ensure a fair and open process.

6.4. Implications for Journals and Publishers

At the journal and publisher level, DtD recommends AI policies go beyond disclosure without abandoning it. Current policies clarify that AI cannot assume authorship, and humans stay responsible for AI-assisted work^{[5][2][6]}. The challenge is ensuring human accountability goes beyond mere declaration.

One approach is to differentiate between assistance disclosure and responsibility affirmation. Assistance disclosure highlights relevant technological contributions, while responsibility affirmation ensures that authors have critically appraised the material, claim responsibility for it, and stand behind it. This second element doesn't always need a separate process in each submission; it can be included within authorship declarations or existing research integrity statements.

Journals should avoid turning these affirmations into broad assessments of an author's overall competence. The goal is to clarify the epistemic significance of authorship, not to add another compliance step. Extra verification should only be used when significant scholarly issues cannot be addressed through standard editorial procedures.

6.5. From AI Policing to Authorship Assurance

The broader institutional implication is a shift from AI policing to authorship assurance. AI policing begins by asking whether a prohibited or undisclosed technology was used. Authorship assurance begins with a different question: whether the people receiving scholarly credit can still assume substantive responsibility for the contribution attributed to them.

This matters because technological provenance and epistemic authorship are not equivalent^[25]. Identifying AI-generated text does not establish that human intellectual agency has been displaced;

failing to identify it does not establish that substantive authorship has been preserved. A governance system centred on detection risks targeting the wrong object.

Authorship assurance is better understood as a layered process. In most cases, ordinary mechanisms—peer review, meaningful contributorship declarations, source and data transparency, methodological documentation, and author responses to substantive questions—provide sufficient assurance. Only where material inconsistencies remain unresolved does additional clarification become appropriate. This treats DtD as a principle underlying scholarly responsibility, not a presumption that every author has something to prove.

The long-term task for scholarly publishing is therefore not to preserve an artificial boundary between purely human and AI-assisted research — a boundary likely to become harder to define and less informative as technological mediation grows more sophisticated. The more durable task is preserving the relationship between scholarly credit, human intellectual agency, and substantive responsibility. DtD clarifies what should remain answerable even as the processes through which scholarship is produced continue to change.

6.6. Practical Application of the Duty to Defend

DtD is proposed primarily as a normative principle rather than a standardised verification procedure, and its practical value depends on whether it can inform scholarly publishing without creating disproportionate burdens or encouraging routine suspicion of authors. A proportionate application integrates DtD into existing editorial processes rather than establishing a separate examination system, as Table 1 illustrates: additional scrutiny increases only where substantive scholarly concerns arise and remain unresolved.

At the ordinary assurance level, no additional authorship verification is normally necessary. Peer review, contributorship declarations, AI-use disclosure where applicable, source and data transparency, and ordinary author responses remain the principal mechanisms for demonstrating scholarly responsibility; named authors should be presumed to satisfy authorship responsibilities unless substantive reasons arise to question that relationship.

Where a specific inconsistency emerges, targeted clarification may be appropriate: an editor might ask an author to explain a central methodological decision, reconcile contradictory interpretations, identify the evidential basis for an important claim, clarify an apparent source problem, or explain the relationship

between declared contributions and a disputed part of the manuscript. Such questions should address the substantive concern, not attempt to determine whether particular language was AI-generated.

Only where significant concerns remain unresolved should enhanced assurance be considered, requesting relevant research records, analytical documentation, source materials, contributor clarification, or a more detailed account of the reasoning behind the disputed contribution. Direct discussion with authors may be appropriate in exceptional cases but should not become a default requirement or be treated as an examination of general competence.

Assurance level	When applicable	Appropriate response
Ordinary	No substantive concern	Normal peer review, disclosure, and contributorship processes
Targeted	Specific scholarly inconsistency	Focused clarification of claims, methods, evidence, or contribution
Enhanced	Significant unresolved concern	Supporting records, contributor clarification, or direct author engagement

Table 1. Practical Application of the Duty to Defend

Note. DtD operates as a proportionate authorship-assurance principle, not a mandatory verification sequence. Additional scrutiny should arise only from substantive scholarly concerns.

As Table 1 shows, ordinary publishing processes are usually sufficient. Targeted assurance is needed only when concerns justify it. All levels should be proportionate, contribution-sensitive, and fair. The goal is to assure authorship, confirming the author has enough epistemic command to be responsible, not to identify who wrote specific words.

6.6.1. An Illustrative Case

A brief illustration shows how the three assurance levels in Table 1 operate in practice. Consider a manuscript proposing a novel theoretical framework in a social-science journal, where the authors disclose that generative AI assisted with literature synthesis and prose refinement.

At the ordinary level, disclosure is complete, the argument is internally coherent, and reviewers raise only standard substantive critiques of the kind any manuscript might receive. No authorship-specific concern has arisen, so peer review proceeds as usual and DtD imposes no additional requirement.

Suppose instead a reviewer notices the manuscript attributes a specific causal mechanism to a cited source that, on inspection, does not support the claim, and that a parallel inconsistency appears in how two sections define the same theoretical construct. This is a targeted concern: substantive, localised, and resolvable through ordinary editorial correspondence. The appropriate response is not to ask whether AI generated the passage, but to ask the authors to clarify the evidential basis for the mechanism and reconcile the conflicting definitions. If they can — explaining why the source was invoked, how the discrepancy arose, and which definition reflects their intended argument — the concern is resolved, and no further scrutiny is warranted. A fluent, immediate answer is not required; under section 7.2, authors should have reasonable opportunity to consult their own materials.

Suppose, alternatively, the authors cannot account for the mechanism or explain which definition they intend, and further inspection finds that several central citations do not exist or do not support the claims attached to them — a pattern suggesting citations were generated rather than verified. This moves the case to the enhanced level: no longer a single resolvable clarification, but a pattern indicating the named authors may lack sufficient command of the manuscript's substantive content. Here it becomes appropriate for the editor to request supporting research records, ask co-authors to individually account for their claimed contributions, or, in serious cases, request direct discussion with the authors. Consistent with section 7.4, the burden at this level should still be proportionate to the concern identified, and authors should be given a clear account of what is being asked and why.

Four principles should govern practical application across levels: substantive trigger, proportionality, contribution sensitivity, and procedural fairness. Verification should stem from a scholarly concern; responses should be proportionate; expectations should match the author's actual contribution; and authors should understand the issue and have a fair chance to respond using standard scholarly resources. This positions DtD as authorship assurance, not surveillance. Its goal is not to prove who generated specific words but to maintain a meaningful link between scholarly credit and the human intellectual agency responsible for the contribution.

7. Risks and Limits of Authorship Verification

The Duty to Defend links authorship to responsibility, but applying it in publishing risks encouraging unnecessary scrutiny and suspicion, which can be counterproductive. It's crucial to distinguish between ensuring authorship and policing it. DtD supports the former only if its evidential and procedural limits are acknowledged.

7.1. Defensibility Is Evidence, Not Proof

An author's ability to explain or defend a contribution can provide evidence of epistemic command, but it cannot conclusively establish authorship authenticity^{[4][7]}. Intellectual understanding is not perfectly observable, and any attempt to infer it from particular responses will be incomplete: an author may understand a complex argument but explain it poorly in an unfamiliar setting, while another may give a fluent account without having played the substantive role their authorship implies.

Treat defensibility as evidential rather than forensic. It may contribute to an overall assessment where legitimate concerns have arisen but should not operate as a standalone test capable of proving or disproving authorship; research records, contributor roles, methodological documentation, source verification, revision histories, and the coherence of responses across the editorial process may all provide complementary information.

This matters because DtD concerns a substantive intellectual relationship, not a directly observable event. The framework identifies what authorship normatively requires; it does not claim that epistemic command can always be measured with precision.

7.2. Unequal Capacity to Perform Knowledge

A further risk arises from treating the performance of knowledge as equivalent to possessing it. Authors differ in their ability to respond rapidly, verbally, confidently, or in a particular language, and communication conditions, memory, disciplinary convention, anxiety, disability, cultural expectations, and question format can all affect how intellectual command becomes visible.

The relevant distinction is between performance asymmetry and epistemic asymmetry: the former concerns differences in the ability to display knowledge under particular conditions; the latter, differences in the underlying intellectual relationship to the contribution. DtD concerns the second, but poorly designed verification practices could easily mistake the first for it.

An inability to give an immediate or polished response should not be treated as evidence that an author lacks epistemic command. Where clarification is legitimately required, authors should ordinarily have sufficient opportunity to consider the question and, where appropriate, consult the materials through which scholarly work is normally conducted^[26]. The relevant standard is a reasonable scholarly account, not performance under examination conditions.

7.3. Selective Suspicion and Verification Bias

Authorship verification risks verification bias if some scholars face unjustified scrutiny. Generative AI may worsen this by triggering suspicions based on writing style, linguistic quality, affiliation, geography, discipline, or expected author sound. Such cues are insufficient for increased review. If DtD is used only when a manuscript “sounds AI-generated,” authorship claims might reinforce inequalities and offer limited insight into epistemic responsibility.

A more robust approach requires clear scholarly triggers. These could include unresolved contradictions in an author’s explanation of key methods, serious doubts about cited evidence, unexplained differences in contributor roles, or significant inconsistencies beyond what peer review can resolve. The concern should stem from the scholarship or documented authorship process, not from assumptions about the submitter.

7.4. Editorial Power and Procedural Fairness

Introducing any authorship-verification mechanism also alters the relationship between editors and authors. Editors already exercise considerable authority over publication decisions; DtD should not extend that authority into an unrestricted capacity to test a scholar’s intelligence, memory, linguistic fluency, or general competence.

Procedural fairness therefore requires a clear distinction between evaluating the submitted scholarship and evaluating whether a legitimate authorship concern has been resolved. Where additional clarification is requested, authors should understand the nature of the concern, how the questions relate to it, and the possible implications of their response — and should have reasonable opportunity to provide supporting records or explanations, rather than being judged through isolated exchanges.

The burden of authorship assurance should remain proportionate to the concern: routine manuscript evaluation should remain routine, and more intensive clarification requires stronger justification. DtD does not reverse the presumption that named authors are legitimate authors; it specifies an obligation

that becomes relevant only when substantive scholarly circumstances reasonably call that relationship into question.

7.5. Collaborative and Distributed Authorship

Authorship verification becomes especially challenging in collaborative research because epistemic labour is often unevenly distributed. Contributorship frameworks recognise diverse roles across activities such as conceptualisation, methodology, analysis, data curation, writing, supervision, and more^[8]. More broadly, epistemic collaboration enables knowledge production through distributed expertise, without requiring every participant to master the entire intellectual process^[13].

A single-standard verification method would conflict with the realities of modern scholarship. An author mainly responsible for methodology, for instance, should not be expected to defend a theoretical contribution as thoroughly as the primary conceptual contributor. However, such differentiated expertise does not negate collective responsibility for the overall integrity and claims of the jointly endorsed work.

DtD must be sensitive to individual contributions. The key question is not whether every author can defend every component equally, but whether each understands their specific contribution and has a sufficient grasp of the central work they endorse. This distinction is crucial for ensuring epistemic authorship aligns with legitimate specialisation.

7.6. The Temporal Problem

Epistemic command is also temporally unstable. Authors might be asked about research conducted months or years earlier. Details that were once easily accessible may no longer be immediately retrievable, especially in complex empirical projects with large data sets, code, literature, or collaborative decisions.

DtD should not be construed as a definitive memory obligation. An author's ability to reconstruct reasoning through credible scholarly records may itself exemplify ongoing epistemic responsibility. Notes, data, code, correspondence, analytical documentation, earlier drafts, and other research materials can constitute part of a reasonable defence where immediate recall is impractical. The temporal dimension emphasises the difference between epistemic command and performance. It's not about retaining every detail indefinitely, but whether the author can reasonably reconstruct the necessary intellectual relationship to justify the contribution.

7.7. Avoiding a New Authorship Orthodoxy

DtD should not evolve into a new universal orthodoxy that supplants one simplistic conception of authorship with another. Authorship practices differ across various disciplines, methodological traditions, research teams, and publication cultures. The epistemic expectations for a single-author conceptual article will differ from those for a large experimental collaboration involving highly specialised contributors.

Furthermore, authorship attribution issues are not exclusive to generative artificial intelligence. Ghostwriting and honorary authorship exemplify longstanding instances where formal attribution and substantive contribution may diverge^[9]. While artificial intelligence enhances the technological scalability of these problems and may render them less apparent in the final text, it is important to recognise that authorship assurance should not be approached under the assumption that all current authorship concerns stem solely from AI utilisation.

The Duty to Defend should remain a normative principle rather than a strict verification protocol. Its goal is to clarify what constitutes accepted scholarly authorship while giving disciplines and institutions the flexibility to recognise this principle within their legitimate practices. Implementation should remain proportionate, sensitive to contributions, procedurally fair, and clear in distinguishing evidence of epistemic authority from proof of authorship authenticity.

The right to be called an author hinges on being accountable for the intellectual contribution, not passing an interrogation. Scholarly publishing must uphold this principle without turning authorship into surveillance.

8. Discussion

As generative AI becomes more involved in scholarly work, discussions around disclosure, attribution, integrity, and the boundaries of technological help have intensified. While much debate centres on whether AI was used, the extent of its contribution, and whether to disclose it, this framework proposes a different focus. The key question is about what human authorship means when scholarly work is increasingly shared between humans and non-human agents.

8.1. Authorship as an Epistemic Status

The first implication is that scholarly authorship should be understood not only as a mechanism for allocating credit and accountability, but as an epistemic status. To be identified as an author is to claim a particular relationship with a contribution to knowledge. Existing standards connect authorship with substantive contribution and accountability^[2], while recent scholarship increasingly emphasises reasoning ownership, epistemic responsibility, claim accountability, and human answerability in AI-mediated scholarship^{[3][4][7]}.

Uebel and Hamamra's^[10] account of authorship as relational responsibility provides an important adjacent perspective: by locating authorship within distributed socio-technical systems while preserving human epistemic stewardship and justificatory answerability, it moves beyond exclusive textual production as the basis of authorship. The present framework develops this direction differently. Rather than treating answerability primarily as a feature distinguishing responsible human authorship within distributed production, DtD specifies a reciprocal normative obligation arising from claiming scholarly author status: scholarly credit and recognition carry an obligation to retain sufficient epistemic command to make the responsibility attached to that status meaningfully exercisable.

This reframing prevents authorship from becoming purely administrative: a name attached to a manuscript alongside declarations of contribution, disclosure, and accountability. Such mechanisms remain important, but if authorship is to keep signifying intellectual contribution to knowledge, the relationship between the credited individual and that contribution must remain substantive rather than nominal.

The right to be called an author is therefore inseparable from an epistemic obligation. "Right" here does not imply an unconditional entitlement to author status; it expresses the reciprocal structure developed in this paper that the scholarly recognition attached to authorship carries responsibility for maintaining an adequate intellectual relationship with the contribution being recognised.

8.2. From Provenance to Epistemic Responsibility

A second implication concerns the limits of provenance-based approaches to AI governance. Disclosure remains important for transparency, and questions of when AI use should be disclosed are becoming more differentiated. Resnik and Hosseini^[27], for example, argue that disclosure requirements should depend on the nature and significance of AI's contribution, rather than assuming every use warrants

identical treatment. Such work illustrates the growing complexity of governing AI involvement through disclosure alone.

DtD addresses a different question. Disclosure concerns the visibility of technological assistance; DtD concerns the author's epistemic relationship with the resulting scholarship. A manuscript may pass through multiple stages of human drafting, AI restructuring, human correction, AI-assisted synthesis, and further revision — so identifying the provenance of every sentence may reveal progressively less about who exercised substantive scholarly judgement.

The more consequential question is whether generated or externally mediated material has been incorporated into human judgement. This shifts attention from textual provenance to epistemic responsibility, without making disclosure irrelevant — it clarifies what disclosure cannot establish. Knowing that AI contributed to a manuscript provides information about process but does not by itself establish whether generated claims were understood, critically evaluated, verified, or responsibly adopted.

This also explains why AI detection provides an inadequate foundation for long-term authorship governance. Even reliable identification of AI-generated language would not establish whether human reasoning had been displaced, and the absence of detected AI-generated text would not establish substantive epistemic command. The deeper governance problem concerns the relationship between the author and the scholarship attributed to them.

8.3. Delegation Is Not the Problem; Displaced Authorship Is

The framework also provides a differentiated account of delegation. Scholarship has never been an entirely individual cognitive activity — research depends on collaborators, specialised expertise, databases, instruments, computational tools, editors, research assistants, and other epistemic resources. Distributed knowledge production is therefore compatible with legitimate authorship^[13].

Generative AI complicates this because assistance can extend into activities traditionally associated with substantive intellectual judgement, including interpretation, theoretical synthesis, explanation, and argument construction. However, drawing the boundary between human and machine production would misidentify the problem. The more relevant distinction is the one developed in section 5.3: whether assistance is integrated into human judgement or produces displaced authorship.

Integration allows external assistance to contribute while preserving sufficient human intellectual agency; displacement occurs when substantive intellectual content becomes insufficiently connected to the understanding and judgement of the person claiming authorship^{[4][28]}. The concern is not cognitive offloading per se, but whether reliance on external assistance becomes so extensive that it weakens the critical engagement necessary for epistemic command^[20]. This distinction rejects two symmetrical errors: assuming extensive AI use necessarily undermines authorship, and assuming formal human approval necessarily preserves it.

The framework therefore treats delegation as epistemically consequential based on what happens to human intellectual agency, not the identity of the assisting agent alone—which is why the degree and character of engagement matter more than a binary classification of scholarship as human-written or AI-generated.

8.4. The Manuscript as Evidence of Authorship

A further implication concerns the manuscript's evidentiary role. Scholarly publishing has traditionally inferred a considerable degree of intellectual participation from producing a sophisticated academic work. That inference was never perfect; ghostwriting, honorary authorship, inappropriate attribution, and fragmented collaboration show that the connection between a finished manuscript and the intellectual involvement of every named author has long been imperfect^[9].

Generative AI changes the strength of that inference. High-quality prose, coherent argumentation, and sophisticated presentation can increasingly be produced with forms of assistance that do not require equivalent human intellectual engagement, so textual sophistication becomes a less complete proxy for epistemic authorship.

This does not mean the manuscript should cease to be the primary object of scholarly evaluation, or that authors should routinely face separate verification. Where legitimate substantive concerns arise, other evidence may complement the finished text — reviewer exchanges, contributorship information, research records, methodological documentation, source verification, and targeted clarification can help establish whether formal authorship remains connected to substantive scholarly agency.

This is also why DtD should be understood as an evidential principle rather than an examination regime: the objective is not to replace evaluation of scholarship with evaluation of scholars, but to recognise that the finished text may no longer provide complete evidence of the epistemic relationship authorship conventionally implies.

9. Conclusion

This paper has argued that the central challenge generative AI poses to scholarly authorship is not detection but definition: what must remain human for a named author's responsibility to be more than nominal. Existing publication guidance correctly assigns accountability to human authors regardless of AI involvement, but accountability alone does not establish that the responsibility it assigns can be exercised. The nominal/substantive distinction makes that gap explicit; the Duty to Defend specifies what closing it requires – that an author retain sufficient epistemic command to give a reasonable account of a work's central claims, methods, evidence, and interpretations when legitimately challenged.

This obligation is not a new burden but a consequence of what authorship already is. Claiming scholarly credit and accepting the responsibility it carries are already reciprocal under existing norms; responsibility cannot be exercised if the capacity to exercise it has not been retained. Distributing intellectual labour among human collaborators does not threaten this structure, because responsibility is relocated rather than lost; each element remains under the command of some accountable person. Generative AI cannot occupy that role, since it cannot itself be held answerable: what it produces becomes part of substantive scholarship only once a human author has critically evaluated, verified, and genuinely incorporated it into their own judgement. Absent that, defensibility is not redistributed but simply absent.

In practice, this means editorial scrutiny should track intellectual engagement rather than technological provenance: ordinary review suffices for most manuscripts, targeted clarification addresses specific inconsistencies, and closer scrutiny is reserved for patterns suggesting named authors cannot account for what is attributed to them. The question is never whether AI was used, but whether those receiving credit can stand behind it. The framework has real limits – epistemic command is hard to verify without becoming a burden in itself, the capacity to visibly "perform" understanding is unevenly distributed, and editorial power to demand justification requires safeguards against arbitrary use. These limits do not undermine the argument; they mark the conditions under which it must operate to avoid becoming what it is meant to prevent — a mechanism of suspicion rather than a safeguard for authorship's integrity.

The right to be called an author was never a right to have produced every word of a text. It is the right to be answerable for what that text claims. As the tools available to scholars change, scholarly publishing has the strongest reason to protect that relationship.

References

1. ^{a, b}Rennie D (1997). "When Authorship Fails." *JAMA*. 278(7):579. doi:[10.1001/jama.1997.03550070071041](https://doi.org/10.1001/jama.1997.03550070071041).
2. ^{a, b, c, d, e, f, g, h, i}International Committee of Medical Journal Editors (2026). "ICMJE | Recommendations | Preparing a Manuscript for Submission to a Medical Journal." International Committee of Medical Journal Editors. <https://www.icmje.org/recommendations/browse/artificial-intelligence/ai-use-by-authors.html>.
3. ^{a, b, c, d, e, f}Koo M, Lu MC (2026). "Who Owns the Argument? A Practical Test for Author Responsibility in AI-Assisted Scholarly Writing." *Learn Publ*. 39(3):e2066. doi:[10.1002/leap.2066](https://doi.org/10.1002/leap.2066).
4. ^{a, b, c, d, e, f, g, h}Nascimento DC, da Silva D (2026). "Identical Errors, Different Responsibilities: Epistemic Duty in AI-Assisted Scientific Writing." *AI Ethics*. 6:Article 484. doi:[10.1007/s43681-026-01330-w](https://doi.org/10.1007/s43681-026-01330-w).
5. ^{a, b, c, d, e, f, g, h}COPE Council (2024). "Authorship and AI Tools." Committee on Publication Ethics. doi:[10.24318/ccvrzbms](https://doi.org/10.24318/ccvrzbms).
6. ^{a, b, c, d, e, f}Zielinski C, Winker MA, Aggarwal R, Ferris LE, Heinemann M, Lapeña JF Jr, Pai SA, Ing E, Citrome L, Alam M, Voight M, Habibzadeh F (2023). "Chatbots, Generative AI, and Scholarly Manuscripts." *Colomb Med*. 54(3):e1015868. doi:[10.25100/cm.v54i3.5868](https://doi.org/10.25100/cm.v54i3.5868).
7. ^{a, b, c, d, e, f, g, h, i}Van Zoonen W, Morgan-Thomas A, Tursunbayeva A (2026). "Beyond AI Disclosure: Claim Accountability and Responsible Research in Scholarly Publishing." *Eur Manag J*. 44(4):550–556. doi:[10.1016/j.emj.2026.06.001](https://doi.org/10.1016/j.emj.2026.06.001).
8. ^{a, b, c, d, e}Brand A, Allen L, Altman M, Hlava M, Scott J (2015). "Beyond Authorship: Attribution, Contribution, Collaboration, and Credit." *Learn Publ*. 28(2):151–155. doi:[10.1087/20150211](https://doi.org/10.1087/20150211).
9. ^{a, b, c, d}Wislar JS, Flanagan A, Fontanarosa PB, DeAngelis CD (2011). "Honorary and Ghost Authorship in High Impact Biomedical Journals: A Cross Sectional Survey." *BMJ*. 343(oct25 1):d6128. doi:[10.1136/bmj.d6128](https://doi.org/10.1136/bmj.d6128).
10. ^{a, b, c, d, e, f, g, h, i}Uebel M, Hamamra B (2026). "Authorship After Generative AI: Distributed Creativity and Relational Responsibility." *Philos Technol*. 39(3). doi:[10.1007/s13347-026-01170-w](https://doi.org/10.1007/s13347-026-01170-w).
11. ^aBorgman CL (2007). *Scholarship in the Digital Age*. The MIT Press. doi:[10.7551/mitpress/7434.001.0001](https://doi.org/10.7551/mitpress/7434.001.0001).
12. ^{a, b}Lloyd D (2025). "Epistemic Responsibility: Toward a Community Standard for Human-AI Collaboration." *Front Artif Intell*. 8:1635691. doi:[10.3389/frai.2025.1635691](https://doi.org/10.3389/frai.2025.1635691).
13. ^{a, b, c, d}Palermos SO (2022). "Responsibility in Epistemic Collaborations: Is It Me, Is It the Group or Are We All to Blame?" *Philos Issues*. 32(1):335–350. doi:[10.1111/phils.12230](https://doi.org/10.1111/phils.12230).
14. ^aThorp HH (2023). "ChatGPT Is Fun, But Not an Author." *Science*. 379(6630):313. doi:[10.1126/science.adg7879](https://doi.org/10.1126/science.adg7879).

15. ^ΔStavropoulos A (2026). "Commentary: Qi, J. et al. Towards Trustworthy Agentic AI: A Comprehensive Survey of Safety, Robustness, Privacy, and System Security. *Acad. AI Appl.* 2026, 2, 8260." *Acad AI Appl.* 2(3). doi:[10.20935/acadai8435](https://doi.org/10.20935/acadai8435).
16. ^ΔSaputra RF, Wibisono MI, Winarno A, Subagyo S (2026). "The Epistemic Authorship Crisis in the Age of Generative AI: Overcoming the Responsibility Gap through Hyper Justification Obligations." *Int J Econ Commerce Manag.* 3(1):18–23. doi:[10.62951/ijecm.v3i1.1090](https://doi.org/10.62951/ijecm.v3i1.1090).
17. ^ΔBiagioli M (1998). "The Instability of Authorship: Credit and Responsibility in Contemporary Biomedicine." *FASEB J.* 12(1):3–16. doi:[10.1096/fasebj.12.1.3](https://doi.org/10.1096/fasebj.12.1.3).
18. ^ΔPatel L, La Vecchia C (2024). "Improving Nutritional Epidemiology to Take on Global Challenges." *Acad Nutr Diet.* 1(1). doi:[10.20935/acadnutr7441](https://doi.org/10.20935/acadnutr7441).
19. ^ΔIsaac JC (2015). "Varieties of Empiricism in Political Science." *Perspect Polit.* 13(4):929–934. doi:[10.1017/s1537592715002236](https://doi.org/10.1017/s1537592715002236).
20. ^a, ^bGerlich M (2025). "AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking." *Societies.* 15(1):6. doi:[10.3390/soc15010006](https://doi.org/10.3390/soc15010006).
21. ^ΔFormosa P, Banks S, Matulionyte R, Ghasemi O (2024). "Can ChatGPT be an Author? Generative AI Creative Writing Assistance and Perceptions of Authorship, Creatorship, Responsibility, and Disclosure." *AI Soc.* 40(5):3405–3417. doi:[10.1007/s00146-024-02081-0](https://doi.org/10.1007/s00146-024-02081-0).
22. ^ΔBorenstein J, Shamoo AE (2015). "Rethinking Authorship in the Era of Collaborative Research." *Account Res.* 22(5):267–283. doi:[10.1080/08989621.2014.968277](https://doi.org/10.1080/08989621.2014.968277).
23. ^ΔResnik DB, Hosseini M, Hauswald R (2026). "Autonomous Artificial Intelligence, Scientific Research, and Human Values." *AI Ethics.* 6(1). doi:[10.1007/s43681-025-00908-0](https://doi.org/10.1007/s43681-025-00908-0).
24. ^ΔResnik DB, Hosseini M, Pennock RT (in press). "Authorship Responsibilities of Scientists Who Use AI Research Agents." *Front Genet.* <https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2026.1954843/abstract>.
25. ^ΔMeyman E (2026). "Provenance Does Not Decide Authorship: What Recorded AI Involvement Does and Does Not Establish." *Zenodo.* doi:[10.5281/zenodo.21890924](https://doi.org/10.5281/zenodo.21890924).
26. ^ΔStoyanov K (2026). "Ethical LLM-Assisted Research: A Framework for Responsible Delegation, Verification, and Epistemic Value." *arXiv.* doi:[10.48550/arxiv.2608.23644](https://doi.org/10.48550/arxiv.2608.23644).
27. ^ΔResnik DB, Hosseini M (2025). "Disclosing Artificial Intelligence Use in Scientific Research and Publication: When Should Disclosure Be Mandatory, Optional, or Unnecessary?" *Account Res.* 33(2):2481949. doi:[10.1080/08989621.2025.2481949](https://doi.org/10.1080/08989621.2025.2481949).

28. [^]Zakhari R (2026). "Authorship, Moral Responsibility, and Generative AI in Nursing." *Nurs Ethics*. 9697330261465739. doi:[10.1177/09697330261465739](https://doi.org/10.1177/09697330261465739).

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.