DOI: 10.2478/doc-2026-0006

Jagat Bahadur Kunwar

School of Business and Economics, Åbo Akademi University, Turku, Finland

jagat.bahadurkunwar@abo.fi

ORCID ID: 0000-0002-2605-2760

Authenticity and Authorship Beyond the Algorithm: Action, Institution, and Epistemic Responsibility under Generative AI

Article history:

Received	**23 March 2026**
Revised	**6 May 2026**
Accepted	**7 May 2026**
Available online	**23 June 2026**

Abstract: This paper examines whether generative artificial intelligence can be considered an authentic author. While large language models (LLMs) demonstrate advanced linguistic competence and produce coherent academic discourse, they foreground unresolved questions concerning the ontology of authorship and the status of authenticity in scholarly authorship. The article is concerned with authorship rather than writing as such; writing is treated as the practical and institutional site through which authorship becomes visible, contested, and assigned.

Rather than treating authorship as a property of textual output, the paper conceptualises it as an activity across three intersecting perspectives: praxis, institutional recognition, and epistemic responsibility. Extending Hannah Arendt's The Human Condition, authorship is theorised in terms of linguistic labour, linguistic work, and linguistic action, with authenticity located in linguistic action as public, accountable, position-taking engagement. A Foucauldian perspective reframes authorship as an institutional function governed by discursive norms, while the epistemic perspective examines whether LLMs satisfy conditions such as semantic understanding, grounding, and epistemic commitment.

The paper argues that AI can perform linguistic labour, support linguistic work, and approximate institutional markers of authorship, but cannot initiate linguistic action or assume epistemic responsibility.

Keywords: authorship, authenticity, generative AI, large language models, epistemic responsibility

Introduction

Generative AI has made it increasingly difficult to treat authorship as a straightforward relation between a human subject and a written text. Large language models can produce fluent academic prose, reproduce disciplinary conventions, and generate texts that resemble scholarly argumentation. Yet these capacities do not by themselves settle whether such systems can be

authors. The central issue addressed in this article is not simply whether AI can write, but whether it can occupy the authorial position of attribution, responsibility, and answerability.

The classical benchmark for operationalising intelligence in an artificial system has been the Turing Test. In his "many minds" argument, Turing states that, since we cannot necessarily attribute intelligence through direct inward access even in other humans, it is unfair to deny it to machine systems on that basis. Hence, if the performance of a machine system is indistinguishable from that of a human through behavioural indicators such as those set up in the classical Turing Test, then the machine has performed indistinguishably from a human participant (Turing, 1950).

While the Turing Test was proposed in 1950, recent developments in artificial intelligence, especially generative AI such as various versions of ChatGPT, perform remarkably well on this criterion. For example, it has been shown that GPT-4 passes an interactive two-player Turing Test and was judged to be human 54% of the time (Jones & Bergen, 2024). Originally, Turing suggested that a system could be considered intelligent if it fooled human judges about 30% of the time in short conversations. More proprietary and presumably more advanced models have been introduced thereafter. Similarly, in evaluations of the Turing Test over 50 years, despite the inherent controversies surrounding it, it has been concluded that, while machines may not perfectly simulate human cognition, they possess sufficient cognitive capacities to pose ethical dilemmas in their use (French, 2000). For the purposes of this article, however, behavioural indistinguishability is not treated as sufficient for authorship. It merely sharpens the question of whether fluent, human-like textual performance can ground authorial responsibility.

This article is therefore concerned primarily with authorship rather than writing as such. Writing is treated here as the practical and institutional site through which authorship becomes visible, contestable, and assignable under conditions of generative AI. The article does not offer a general theory of academic writing or a comprehensive intervention into writing studies. Instead, it asks what distinguishes AI-assisted textual production from authentic authorship.

Recent debates on AI understanding, agency, and epistemic responsibility have moved beyond the Turing Test, acknowledging machine linguistic performance while asking whether symbolic manipulation in formal systems can amount to semantic understanding (Searle, 1980), whether computers can have subjective experience (Chalmers, 1995), how the architecture of LLMs relates to cognitive capability (Millière & Buckner, 2024; 2025), and what ethical problems follow from the use of generative AI in knowledge production (Hicks et al., 2024). These debates matter for authorship because authentic authorship, as understood here, requires more than behavioural indistinguishability or fluent textual output; it requires some relation to claims, reasons, evidence, commitments, and consequences.

Concurrently with these developments in debates on AI understanding, recent developments regarding authorship, still mostly confined to literary criticism, chart the historical evolution of the concept of authorship and conclude that authorship is primarily an affair of public rather than private consciousness, where the public domain consists of an "objective reality renderable by language, a shared vision of how the social should be restructured, or in terms of public conventions and traditions for the production and reception of discourse" (Burke, 1995, p. xviii). This observation also aligns with the "death of the author" argument proposed by Roland Barthes (1967), wherein the originality of the text is no longer attributed to a singular originating subject but understood as a blending and clashing of prior writings, themselves not original (Barthes, 1967). In this sense, the conventional figure of the author is displaced by the author as a "function of discourse", as articulated by Michel Foucault (1969), a shift that has subsequently shaped debates on authorship, particularly around questions of authorial intent, originality, autonomy, and aesthetic engagement (Puşcaşu, 2024). This is why the present article treats authorship not as mere textual production, but as a position constituted through public attribution, institutional recognition, and responsibility for claims.

These debates become newly consequential when LLMs can approximate many of the visible markers conventionally associated with authorship. At one level, the authorial enterprise is institutionalised through review processes,

templates, argumentative structures, citation practices, genre expectations, and disciplinary norms. These components are not simply matters of private consciousness but of the public and institutional domain. At another level, questions of authorial intent, originality, autonomy, and engagement bring us to questions of AI agency, emergent learning, understanding, and subjective experience. Generative AI therefore complicates not only writing, but also the institutional and epistemic conditions through which authorship is assigned, recognised, and made accountable.

Recent public instances of AI-generated work being submitted to or published in academic venues (e.g., Sakana.ai, 2026) sharpen the problem of authorial attribution. They suggest not simply that AI can generate academic-looking text, but that institutional markers of authorship may sometimes be approximated without the presence of a human authorial subject in the stronger sense of responsibility and answerability. More importantly, it may be the case that even human authors operate within institutionalised and disciplined modes of thinking that are, at least in part, imitable. As such, our creative enterprises may often be routine derivations of insights from given premises rather than sparks of human genius.

For example, Turing's reply to Lady Lovelace's objection, namely the objection that machines cannot originate anything and can only do what they are instructed to do, is instructive here. Turing asks "who could be certain that original work that he has done was not simply the growth of the seed planted in him by teaching, or the effect of well-known general principles" (Turing, 1950). For Turing, the Lovelace objection is weak because human originality itself may not be the wholly non-algorithmic enterprise we often assume it to be. Originality, rather than being fundamentally disruptive, may consist of a series of inductions combined with vast networks of information leading to surprising novelty. If this is so, then it raises the question: what kind of activity is authorship, what role do human authors play in this authorial enterprise, and how much can be safely attributed to LLMs?

These developments raise the central question of this article: can AI be regarded as an authentic author? The question is not whether AI can produce academic prose, but whether it can occupy the authorial position of responsibility,

attribution, and answerability. To address this question, the paper develops three complementary perspectives on authorship: praxis, institutional, and epistemic.

The first perspective draws on Hannah Arendt's distinction between labour, work, and action in The Human Condition (Arendt, 1958). Labour refers to repetitive, necessary, and consumptive activity; work refers to the production of relatively durable artefacts; and action refers to speech and deed through which persons appear before others, disclose who they are, and become answerable in a shared world. Translated into the present argument, generative AI can perform linguistic labour, such as drafting, summarising, paraphrasing, and stylistic smoothing. It can also support linguistic work, such as the production and organisation of durable scholarly artefacts. However, it cannot itself perform linguistic action, because it does not appear as an embodied and accountable subject before others.

The praxis perspective asks what kind of human activity authorship is and where, if anywhere, authenticity resides. The institutional perspective asks how authorship is authorised, attributed, and regulated within scholarly discourse. The epistemic perspective asks whether current LLMs can satisfy conditions associated with authorship, including semantic understanding, grounding, agential stability, and epistemic commitment.

The article addresses debates on AI authorship, scholarly attribution, research integrity, and epistemic responsibility in AI-mediated knowledge production. By integrating these three perspectives, the paper contributes to debates on AI authorship by clarifying why fluent textual production is not equivalent to authentic authorship. As will be apparent from the following sections, recent developments in AI collapse commonly understood proxies for voice, such as prose quality, forcing a re-examination of the authorial enterprise, authorial voice, and authenticity in scholarly discourse. Similarly, to the knowledge of the author(s), Arendt's ontology of human action has not been extended to authorial activity in this specific context. This paper therefore uses Arendt's labour–work–action distinction to differentiate linguistic labour, linguistic work, and linguistic action, thereby making it more practical to distinguish between AI and human roles in the authorial enterprise.

Additionally, the paper proposes a schematic framework for understanding authorship from three complementary but multidisciplinary perspectives: the praxis perspective, the institutionalised perspective, and the epistemic perspective. The framework positions authenticity in authorship at the nexus of public responsibility, institutional recognition, and epistemic standards. Building on these arguments, we distil a set of characteristic features that, together, clarify the debate on authenticity in authorship in relation to LLMs.

Furthermore, the paper develops the schematic analytically to distinguish authentic, performative, marginal, and simulated forms of authorship, thereby situating AI-generated and AI-assisted texts within a broader typology of authorial practice. More specifically, the contribution lies in integrating an Arendtian ontology of human activity with an institutional and epistemic account of authorship, and in developing a schematic that not only conceptualises these dimensions but also analytically differentiates forms of authorship and locates AI-assisted textual production within this structure.

The paper is structured as follows. First, it develops an Arendtian account of authorship as linguistic labour, linguistic work, and linguistic action. Second, it examines authorship as an institutional function through which scholarly authority, attribution, and accountability are organised. Third, it considers relevant debates on AI understanding, grounding, subjective experience, and epistemic commitment. The final section integrates these perspectives into a schematic account of authentic authorship under conditions of generative AI and briefly outlines directions for future research.

Authorship as praxis: linguistic labour, work, and action

Referring to Hannah Arendt's The Human Condition (Arendt, 1958), which distinguishes the ontology of human activity into labour, work, and action, we may conceptualise authorship as a human activity rather than merely the production of text. Arendt refers to labour as human activity comprising activities that are repetitive, necessary, and consumptive. Work, as opposed to

labour, is the production of durable artefacts, dealing with objectification and world-building activities. Lastly, action comprises activities that necessitate the disclosure of the "who", plurality, unpredictability, and the publicness of the enterprise. This section therefore treats writing only insofar as it materialises different forms of authorial activity.

If we recast the labour–work–action distinction in relation to authorship, these components become visible in the practices through which scholarly texts are produced. For example, much of the authorial enterprise, such as drafting boilerplate, summarising literature, reproducing genre norms, and iterative optimisation of the draft for clarity and fluency, including smoothing prose, paraphrasing, and summarising, can be considered "linguistic labour" because, as per Arendt's definition (Arendt, 1958; Saeidnia & Lang, 2017), these activities are repetitive, rule-governed, and oriented towards efficiency and academic consumption. It is important to note that AI performs linguistic labour extraordinarily well and is structurally strong precisely for these reasons. Linguistic labour, in this sense, does not by itself ground authentic authorship, because labour is not where authenticity resides in Arendt's account.

If we think about the conceptualisation of "work" in Arendt's sense, then we can say that "linguistic work" is about producing durable textual artefacts such as articles and reviews, stabilising arguments into objects, creating conceptual architecture, methodological inscriptions, and overall contributing to the archive of scholarship. In essence, linguistic work may be defined as scholarly artefact--making. AI also assists here, but some components of linguistic work still require human purposive restructuring and evaluative judgement, perhaps in the form of iterative prompts. It may be suggested that academic work is not always oriented towards action in the Arendtian sense, and the role of AI becomes problematic primarily when work is conflated with action. The point is therefore not to classify all writing as authentic or inauthentic, but to distinguish the forms of authorial activity that AI can assist from those it cannot occupy.

If we further use Arendt's notion of action (Arendt, 1958; Saeidnia & Lang, 2017) and define "linguistic action" as something that involves taking responsibility for a position, inserting oneself as an author into a contested public space, such as conferences and seminars, and risking disagreement,

reinterpretation, and response, then, in this sense, authorial voice is more than style. Linguistic action is a public, risky, identity-disclosing intervention into a plural space, taking responsibility for a position and initiating contestation. Voice is therefore about answerability and not merely fluency or prose quality. Linguistic action involves appearing as an embodied actor before others, presupposes plurality, and requires intelligibility to others. In this regard, AI cannot act in the true Arendtian sense because it cannot be answerable in the human public realm. Therefore, authorial voice is not merely a stylistic signature, but rather the residue of actions that involve accountability, risk, and positioning, even within linguistic work and labour.

Based on the labour–work–action distinction, and judging by the current capabilities of AI, we can claim that AI can perform linguistic labour remarkably well and assist greatly, or even fully, in linguistic work, but cannot initiate linguistic action. Only humans are able to do that by assuming public responsibility for meaning. We can conclude at this stage that when we talk about authorial authenticity or voice, it denotes responsibility for one's claims, errors, and explicit positioning, rather than merely a recognisable textual style. An author is responsible for the claims they make and the errors they commit, and is accountable for their explicit positioning, signifying that authentic authorship aligns with Arendtian linguistic action within the labour–work–action distinction. This conceptualisation explicitly resists a reductionist view of authorship as mere output and treats it as an activity.

However, such responsibility is not exercised in isolation but presupposes a shared epistemic space within which claims can be evaluated and contested. To add to the above observation, Chakrabarty's (1998) concern with the limits of epistemic pluralism highlights a condition often overlooked in debates about authenticity and voice. Albeit in the context of minority and subaltern histories, while they rightly contest dominant narratives, Chakrabarty warns that the rejection of shared criteria of fact and evidence risks undermining the very possibility of adjudicating competing claims in public life. This insight complements an Arendtian understanding of action as appearing before others in a shared world: without minimal agreement on what counts as evidence, speech ceases to be action and becomes either private expression or irreconcilable

assertion. Authenticity, therefore, cannot be equated with unrestricted self--expression; it presupposes participation in a common epistemic world that renders claims contestable and answerable. This also clarifies that truth is not treated here as privately possessed or politically neutral, but as publicly contestable within shared epistemic practices.

Authorship as an institutional function

In academic publishing, authorship is not merely a description of who produced a text. It is an institutional mechanism through which credibility, authority, ownership, responsibility, and eligibility to speak are distributed. Journals, peer review, citation practices, affiliation, editorial judgement, authorship guidelines, and disciplinary conventions help determine whose claims become recognisable as scholarly contributions and whose voices remain peripheral or unauthorised (Cronin, 2001; Chartier, 2003; Hyland, 2009; COPE Council, 2019). In this sense, much of the ontology of authorship is not merely psychological but procedural. Examples include the peer review process, citation practices, journal styles, and accountability mechanisms for authenticity and originality. Often, it is institutions that define what counts as meaningful, reproducible, and credible speech. The preoccupation of an author with analysing various journal metrics, such as citation density, genre conventions, and reviewer expectations, is already part of these authenticity-producing mechanisms. If we look closely, it becomes apparent that AI does not "kill authenticity" but exposes how little authenticity may be required for institutional recognition.

Similarly, as authors we are involved in the enterprise of reproducing or stabilising discourses, for example, the literature review in this paper, and, to some extent, challenging the discourse. If we consider LLMs, they are already capable of sedimenting past discourses and reproducing statistical crystallisations of academic norms, because they are trained on large corpora that include academic and academic-like texts. Effectively, the training data in LLMs already encode prior regimes of truth.

Following Foucault on discourse and the author-function, we may consider an author not simply as a person but as a regulatory node where sedimented discourse and crystallised academic norms precede and exceed the individual writer. The author-function is embedded within legal and institutional arrangements that delimit and organise the field of discourse. Its operation is historically and culturally variable, rather than stable or universal across contexts. It does not arise from a straightforward attribution of a text to an originating subject, but is constituted through a set of structured and often implicit procedures that govern how authorship is assigned. As such, it cannot be reduced to, or equated with, a concrete individual (Foucault, 1969; Chartier, 2003; Pușcașu, 2024).

If, then, authenticity is institutionally defined, and if AI-produced texts largely satisfy those institutional criteria, the objection to AI on the grounds of inauthenticity is partly jurisdictional rather than purely ontological: it concerns who is authorised to occupy the position of a speaking subject, who may occupy the author-function, and who can be held answerable. In other words, the question concerns authority and responsibility as much as textual production.

For the purposes of this article, the key point is that a Foucauldian understanding of authorship shifts authenticity away from personal sincerity and towards authorisation, recognition, and responsibility within institutional procedures. A systematic analysis of authorship as an institutional function therefore requires attention to the author-function, regimes of truth, and discursive authority, rather than to personal sincerity alone. The concern with AI is that it already incorporates many of the procedural logics through which authenticity is typically assessed, thereby blurring, and in some cases dissolving, the distinction between the authentic and the non-authentic at the level of institutional recognition.

This shift away from authenticity as interior sincerity towards its procedural and publicly recognisable dimensions also opens onto a complementary, non-interiorist account of authenticity. In this regard, Charles Guignon (2004) reorients the debate away from romantic self-expression and the metaphysics of psychological depth, and instead understands authenticity as situated,

practice-bound, and answerable to shared norms. From this perspective, authenticity does not require that a text be produced "from inside" a sovereign self, but can obtain without any strong commitment to interiorism. Rather, authenticity is practical, relational, and institutionally situated, not solely private or reducible to institutional approval, but concerns how an author inhabits norms without being wholly subsumed by them. Authentic scholarly voice is therefore not a distinctive style, but a pattern of responsible self-ownership in making claims, where the author can give reasons for their position, acknowledge the standards they are using, and stand behind the consequences of those commitments within a community of inquiry.

Bringing in the Foucauldian perspective on the author-function (Foucault, 1969; Chartier, 2003; Pușcașu, 2024) and Guignon's view of authenticity (Guignon, 2004) as practice-bound rather than a metaphysical property, we can conceptualise authorship not as an inner property of a subject, the author in this case, but as a discursively authorised classification. In other words, authenticity is recognised rather than expressed, voice is not possessed but validated, and originality is, by definition, institutionally ratified rather than grounded in private consciousness. Thus, the institutions that define authenticity may include editors, reviewers, disciplines, and other gatekeepers in the publication process.

This discussion suggests that authorship is partly an institutional function, extending to questions of attribution, accountability, credit, and disclosure. If authorship is an institutional function and is deeply involved in discourse reproduction, AI can act as an amplifier of genre norms and stabilise disciplinary style to a greater extent. However, once again, AI is limited in terms of epistemic responsibility when it comes to the accountability of the author regarding claims, errors, and positioning.

The point, closely related to the Arendtian analysis, is that debates about authenticity often misidentify linguistic labour or linguistic work as linguistic action. It is important to note here that we subscribe neither to total institutional determinism in the Foucauldian sense nor to the naïve humanism that might be suggested by Arendt alone in understanding authenticity. Both perspectives are required, as authorship is neither totally institutionally determined nor a heroic labour in which authentic voice emerges from expressive originality

alone. Whereas Foucault shows who defines authenticity in the sense of what makes a claim admissible as authentic knowledge, Arendt identifies what kind of linguistic activity is at stake, albeit without specifying epistemic criteria for intelligibility.

Epistemic perspectives on AI and capabilities

The final perspective concerns AI understanding, grounding, agency, and epistemic commitment. These debates matter for authorship because authentic authorship, as defined here, requires more than the production of fluent text. It requires some relation to meaning, reasons, evidence, claims, and consequences. The first issue concerns whether current LLMs "understand" language, a notion for which there is no settled or widely agreed meaning, especially given that authorship is primarily a linguistic enterprise. The second issue concerns whether LLMs can have "subjective experience", again a concept marked by competing interpretations and no definitive consensus, as our definition of linguistic action assumes accountability in public as a marker of authenticity, which in turn presupposes some form of agency and epistemic norms derived from authorial subjectivity.

On "understanding" language

The issue of understanding matters here only insofar as it bears on authorship: fluent linguistic production does not automatically entail authorial understanding, commitment, or responsibility. We must first clarify what is meant by "understanding language". In human cases, understanding presupposes an embodied speaker, intentions behind speech acts, and the ability to relate linguistic expressions to causal realities. When we say that LLMs do not understand language, we are making a claim that goes beyond mere cognitive competence: current LLM behaviour does not amount to the kind of understanding humans exhibit. The key distinction is between formal linguistic competence and pragmatic understanding. Producing well-formed

or convincing sentences is not sufficient; understanding involves grasping the meaning of those sentences in relation to the conditions from which they arise.

Our position aligns with the distinction between syntactic performance and semantic understanding. Searle's Chinese Room argument (Searle, 1980) shows how a system can generate appropriate linguistic outputs by manipulating symbols according to rules without grasping their meaning. The argument imagines a person manipulating Chinese symbols according to formal rules without understanding Chinese, thereby showing how appropriate linguistic output may be produced without semantic comprehension. From the outside, the system appears to understand the language, but internally it performs only formal operations. Applied to contemporary LLMs, fluent and contextually appropriate responses may therefore reflect syntactic pattern manipulation rather than genuine semantic comprehension. Related critiques, such as the blockhead argument and the stochastic parrot argument (Bender et al., 2021; Millière & Buckner, 2024), similarly suggest that systems trained to predict sequences of tokens can reproduce convincing linguistic outputs without possessing human-like understanding. If authorship is understood as accountable activity rather than textual output, this distinction is crucial.

Rejecting Turing's equivalence between intelligence and behavioural performance (Turing, 1950), Searle argues that genuine understanding requires intentionality, that is, mental states directed towards objects or states of affairs (Searle, 1980). LLMs manipulate symbols according to formal rules without attaching meaning to them. While such systems may generate appropriate outputs, the process remains purely syntactic. Understanding, by contrast, requires semantic relations to the world. The Chinese Room argument therefore suggests that rule-governed symbol manipulation alone cannot produce genuine understanding.

Understanding may also be linked to embodiment, where meaning emerges through sensorimotor interaction with the world. For authorship, embodiment matters because claims are made from situated positions and can be answered, challenged, or sanctioned within a shared world. Text-only systems lack such embodied cognition (Searle, 1980). This concern connects to the broader issue

of grounding. Understanding requires linguistic symbols to be connected to entities and causal relations in the world. Although LLMs show strong syntactic performance and may encode inferential relations between expressions, this does not guarantee referential meaning. Trained primarily on textual data rather than interaction with the world, any world models they possess emerge indirectly from linguistic patterns. As a result, LLMs may exhibit inferential semantic competence while lacking grounding, referential semantic competence, and robust representations of causal structure (Millière & Buckner, 2025). Describing model errors as "hallucinations" can therefore be misleading. Hallucination implies misperception, which risks anthropomorphising the system (Kang & Shaw, 2024), whereas LLMs do not perceive the world at all. Their outputs arise from probabilistic pattern generation rather than misperception, reflecting the same underlying limitations of grounding, reference, and causal connection to the world (Hicks et al., 2024). This is especially relevant where authorship involves situated forms of knowledge, since the author's relation to a world, community, or position can shape what is being claimed and what is at stake in making the claim.

Human language use is embedded in communicative intentions and broader patterns of action and social interaction. The meaning of linguistic expressions often depends on a speaker's intention within a communicative act, where language is used to achieve goals such as coordinating activity, expressing beliefs, or sharing information about the world (Millière & Buckner, 2025). By contrast, current LLMs are optimised for linguistic plausibility (Hicks et al., 2024), generating plausible continuations of text rather than participating in communicative action. Although their outputs may appear syntactically and semantically convincing, they are not grounded in genuine communicative intentions or practical engagement with the world, which raises doubts about whether they reflect genuine language understanding. If linguistic action is the hallmark of authentic authorship, as outlined previously, LLMs, devoid of communicative intentions, cannot necessarily participate in linguistic action.

Human communication presupposes agents with relatively stable beliefs, viewpoints, and commitments over time. This stability helps maintain consistent meanings in dialogue. By contrast, LLMs readily change positions when

prompted and can be induced to defend contradictory claims. This malleability suggests that their outputs are not anchored in a stable agent perspective, which undermines the idea that their utterances carry determinate meanings in the way human linguistic expressions do (Millière & Buckner, 2025). An authentic author probably has some degree of agential stability over time, which the current generation of LLMs lacks (Hicks et al., 2024).

This observation also connects to debates on the conditions under which something may be regarded as an autonomous agent. While some accounts attribute a form of instrumental agency to LLMs (Yildirim & Paul, 2023), others argue that such systems cannot be said to know anything precisely because they lack agency (Goddu et al., 2024). For the present argument, however, the key issue is narrower: whether LLMs can occupy the authorial position of responsibility, commitment, and answerability. Related debates on AI agency and free will remain unsettled, but what is at stake here is not LLMs as such, but how agency and knowledge are conceptualised in relation to authorship.

Furthermore, human language develops through cultural ratcheting, where speakers build on and stabilise the knowledge and practices of previous generations. This process involves not only producing linguistic outputs but also interpreting, evaluating, and integrating them within a shared cultural context. Although LLMs can generate sophisticated responses, they do not participate in this process of cultural transmission. Humans remain responsible for interpreting and stabilising their outputs, which suggests that LLMs do not meaningfully participate in cultural learning (Millière & Buckner, 2025). If we assume that, as authors, we participate in shared meaning-making within a given cultural context, LLMs could not be considered authentic authors.

Human linguistic assertions operate under norms of truth and epistemic commitment, where speakers present statements as representing how things are in the world. By contrast, LLMs generate sentences without any intrinsic orientation towards truth. This does not imply that human authors possess truth in an unmediated or politically neutral way. Scholarly truth is relationally produced through evidence, interpretation, citation, peer review, contestation, and correction. Precisely because truth claims are socially and institutionally mediated, authorship matters: claims require accountable subjects who can

be questioned, challenged, corrected, and held responsible. LLMs, by contrast, produce text that is statistically plausible given the preceding context. Consequently, when LLM outputs are correct, this accuracy often arises as a statistical by-product of patterns in training data rather than from an attempt to track or represent facts about the world. This truth-indifference distinguishes linguistic plausibility from genuine assertion. Although LLMs exhibit strong syntactic fluency and can track inferential relations between sentences, their outputs remain independent of the epistemic norms (Chakrabarty, 1998) that govern human linguistic practice. Linguistic competence in such systems therefore operates at the level of form and inference rather than referential meaning or epistemic commitment to truth. In this respect, their outputs may be characterised as "bullshit", in the sense of being produced without regard for truth (Frankfurt, 2005; Hicks et al., 2024). Although this is a normative commitment, as authors we cannot risk truth-indifference in academic enterprise, and so we may also add epistemic commitment as an ideal characteristic of authenticity.

Another limitation concerns the architecture of current language models. Most LLMs primarily model formal patterns in language rather than the broader cognitive systems that support human linguistic competence. Human language use is embedded in a wider set of psychological capacities, including perception, memory, attention, imagination, social cognition, and long-term planning, which help maintain coherent and systematic behaviour across contexts. Because current LLMs lack such an integrated cognitive architecture, their linguistic performance does not reflect the same kind of systematic understanding found in human language users (Millière & Buckner, 2025). Although LLMs may produce syntactically and semantically plausible texts, the systematic understanding and articulation of a given phenomenon, which is the subject of any academic article, remain limited and questionable.

Even if LLM performance approaches that of human speakers and appears to "understand" language, there are methodological concerns about such evaluations. In many contexts, their responses are indistinguishable from human outputs, which makes behavioural performance an appealing indicator of understanding (Jones & Bergen, 2024). However, evaluation benchmarks can become targets of optimisation, models may benefit from data contamination

because benchmark questions appear in training datasets, and there are deeper concerns about construct validity. Many benchmarks were designed to assess human cognition and rely on assumptions about how humans process language. As a result, strong behavioural competence on these tests does not necessarily establish that the system genuinely grasps meaning (Millière & Buckner, 2025). If this is the case, this point largely softens the debate on authenticity, as the crucial claim is that, without understanding language, one cannot produce authentic authorship.

A strong counter-argument to our position concerns compositionality in language (Millière & Buckner, 2025). Traditionally, linguistic competence has been understood as rule-based, grounded in explicit grammatical structures such as those proposed in theories of Universal Grammar by Noam Chomsky (Chomsky, 1986). However, the success of LLMs challenges this assumption. These systems produce grammatically well-formed and systematically related sentences despite lacking any explicitly encoded symbolic grammar. This suggests that productivity and systematicity in language may emerge statistically from large-scale pattern learning rather than from rule-based cognitive structures. If compositional structure can arise from patterns of linguistic usage alone, then the symbols and rules posited in classical linguistic theories may be descriptive tools rather than fundamental components of human cognition. On this view, strong linguistic competence might not require the kinds of cognitive mechanisms traditionally associated with language understanding, which weakens the claim that LLMs lack genuine linguistic competence. If emergent linguistic acquisition and understanding are possible through compositionality, a still largely debated issue, this can only be verified empirically, and so the position on language understanding may shift in the future, changing the dynamics of the concepts explored in this paper.

Thus, the point of this section is not to settle the philosophy of language or cognition, but to clarify why linguistic fluency alone cannot ground authentic authorship without understanding, grounding, and epistemic answerability.

Subjective experience, embodiment, and authorial position-taking

The relevance of subjective experience to this article is limited but specific. If authentic authorship involves position-taking, embodied exposure, and answerability, then subjective experience matters not as an abstract property of consciousness but as part of the broader question of whether an author can appear before others, be addressed, and bear consequences for claims. By "subjective experience", we refer to first-person phenomenal awareness, or the felt quality of experience (Nagel, 1974; Chalmers, 1995).

Our position is agnostic regarding machine consciousness. Current evidence does not justify attributing subjective experience to artificial systems such as LLMs, although some philosophical approaches, particularly computational functionalism, leave open the possibility that sufficiently advanced artificial systems could possess conscious experience in principle. The most defensible position is therefore neither to attribute nor categorically deny machine consciousness, but to ask whether current LLMs can satisfy the conditions relevant to authentic authorship.

Turing's discussion of the "argument from consciousness" remains relevant here. The objection claims that a machine cannot truly think because it lacks feelings or subjective experience. Turing responded that this requirement is problematic, since we cannot directly access the experiences of other minds, including those of other humans, and therefore typically infer mental states from behaviour (Turing, 1950). This reply cautions against a simple dismissal of machine consciousness, but it also shows the limits of behavioural evidence: convincing first-person statements generated by an LLM do not by themselves establish subjective experience.

As discussed above, Searle's Chinese Room argument supports the narrower point that formal symbol manipulation does not by itself establish understanding or intentionality, both of which are relevant to authorial responsibility. The discussion of grounding and world models also has implications for authorship. If subjective experience depends partly on embodied interaction with an environment, then purely text-trained systems lack at least some of the conditions associated with situated perspective. This does not settle the question of machine consciousness,

since the relationship between embodiment, cognition, and consciousness remains contested (Chalmers, 2023; Millière & Buckner, 2025). For the present argument, the point is narrower: authentic authorship presupposes an authorial position from which claims can be made, defended, revised, and answered for within a shared world.

Chalmers' (1995) distinction between the "easy" and "hard" problems of consciousness is useful here. The easy problems concern cognitive functions such as perception, information processing, verbal report, attention, and behavioural integration. The hard problem concerns subjective experience itself, or what Nagel (1974) describes as there being "something it is like" to be a conscious organism. This distinction matters because reproducing functional aspects of human communication does not by itself demonstrate that there is anything it is like to be an artificial system (Chalmers, 1995).

These debates remain unresolved. Functionalist accounts suggest that consciousness may depend on patterns of causal or informational organisation rather than biological substrate alone, leaving open the possibility of artificial consciousness in principle. By contrast, biological or embodiment-based accounts remain more sceptical about whether purely artificial or text-trained systems can instantiate subjective experience. For the purposes of this article, this debate need not be settled. The relevant point is that current LLMs do not provide sufficient grounds for attributing the subjective, embodied position--taking presupposed by authentic authorship.

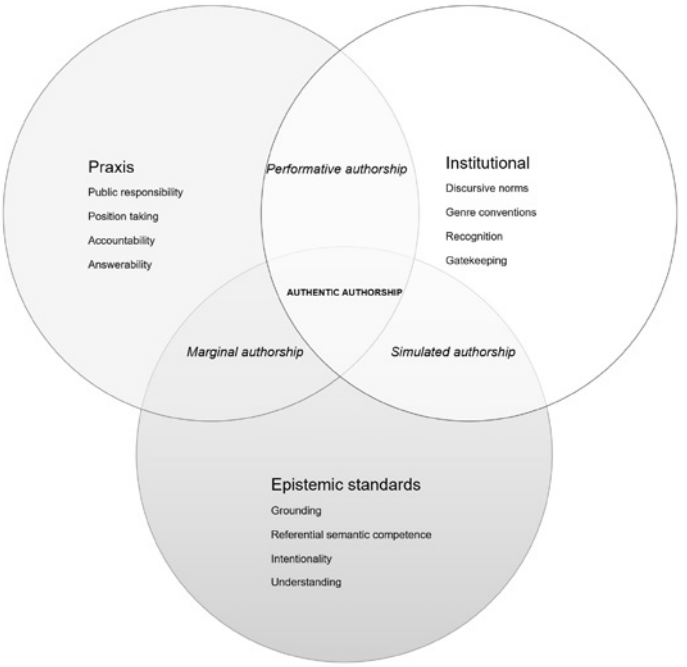
The above discussion illustrates that this remains a developing field and is relevant to authorship only insofar as subjective position-taking implies some form of embodied perspective, reflexivity, and answerability. At present, current evidence does not justify attributing such subjective experience to LLMs. This does not require a categorical denial of future artificial consciousness; it only means that current LLMs cannot be treated as authentic authors on this ground.

The schematic and discussions

By presenting three perspectives on authorship, namely the praxis, institutional, and epistemic perspectives, we can now move towards developing the idea of authenticity in authorship, particularly in relation to the existence and developing capabilities of LLMs. It is important to note here that the view of authenticity derived here is largely normative, as we claim it will always be, but it nonetheless contributes to debates on AI authorship, scholarly attribution, and epistemic responsibility.

Figure 1 shows the schematic representation of authorship as an intersection of praxis, institutional, and epistemic conditions, and illustrates how authenticity emerges at their convergence.

Figure 1. Authentic authorship as the intersection of praxis, institutional recognition, and epistemic responsibility



Source: Author's own elaboration.

While we identified the characteristics of authentic authorship as the nexus of praxis, institutional, and epistemic conditions, this does not indicate a moral hierarchisation of authorial ontology. These are also not three equal, independent, or modular domains, but rather interdependent conditions for authenticity in authorship.

The schematic also allows us to consider the analytical implications of partial intersections among these conditions. Different configurations generate further forms of authorship that clarify the position of AI-mediated textual production within the broader framework.

For example, when we combine the praxis and institutional dimensions, we observe a form of authorship that is recognised and publicly positioned but epistemically weak. An illustrative instance would be work that is institutionally validated yet displays indifference to truth. We label this as performative authorship.

Similarly, at the intersection of praxis and epistemic dimensions, we can infer a form of authorship that is grounded and involves genuine linguistic action, but remains institutionally unrecognised, often situated outside dominant discourse. We label this as marginal authorship, an instance of unrecognised authenticity.

Furthermore, if we combine the institutional and epistemic conditions, we observe a form of authorship that satisfies disciplinary standards of plausibility and institutional criteria, yet lacks the characteristics of praxis such as ownership, responsibility, and position-taking. This form may be described as simulated authorship, representing procedural epistemic production without linguistic action. It is important to clarify here that the epistemic dimension, as represented in the schematic, refers to the appearance of epistemic grounding rather than epistemic grounding in the stronger sense discussed earlier, such as intentionality, referential semantic competence, or genuine understanding of language. While AI systems can reproduce this appearance in terms of plausibility, coherence, and conformity to disciplinary standards, and appear "epistemic" enough to pass institutional filters, they do not satisfy the conditions required for grounding. In this sense, simulated authorship aligns with epistemic form without satisfying epistemic grounding, of which AI-mediated textual production is the most prominent contemporary instance.

In this way, the intersection of all three conditions stabilises authenticity. Rather than positing authorship as a simple dichotomy between human and AI, we conceptualise it as a set of analytical distinctions rather than a moral hierarchy, where different configurations are possible for human authors, while AI-mediated authorship most closely aligns with the intersection of institutional criteria and epistemic plausibility. In this sense, the schematic not only integrates existing theoretical perspectives, but also generates a typology of authorship that clarifies the position of AI within the broader ontology of scholarly practice.

Table 1 disaggregates these dimensions further by showing where current AI systems appear strong, where they provide partial support, and where they remain unable to satisfy the conditions of authentic authorship.

Table 1. Distribution of AI capabilities across the conditions of authentic authorship

Dimension	Authorial condition	AI capability	Authorship implication
Praxis	Linguistic labour	Strong	Does not ground authentic authorship
Praxis	Linguistic work	Partial/supportive	Requires human judgement and organisation
Praxis	Linguistic action	Absent	Requires embodied answerability
Institutional	Genre and procedural compliance	Strong	Can approximate authorial markers
Institutional	Attribution and accountability	Absent/derivative	Requires human or institutional responsibility
Epistemic	Plausibility and coherence	Strong	Can simulate epistemic form
Epistemic	Grounding and epistemic commitment	Weak/absent	Requires responsible authorial judgement

Source: Author's own elaboration.

As Table 1 indicates, AI systems can operate effectively at the level of linguistic labour and can support aspects of linguistic work, while remaining excluded from linguistic action. Similarly, AI can satisfy many forms of institutional procedural compliance, but it cannot assume attributional responsibility or be held accountable for claims. At the epistemic level, AI can reproduce plausibility and coherence, but it does not satisfy the stronger conditions of grounding and epistemic commitment discussed earlier.

The framework summarised in Figure 1 and Table 1 constitutes the central contribution of this paper. While existing debates on AI and authorship often proceed through abstract claims concerning creativity, intention, or intelligence, the present framework disaggregates authorship into analytically distinct yet interrelated dimensions and shows how AI intersects with each unevenly. In doing so, the discussion is shifted away from a binary question of whether AI can be considered an author, towards a more structured account of where and how authorship is distributed, simulated, or enacted within scholarly practice. The contribution of the paper, therefore, is not to resolve the question of AI authorship, but to render it analytically tractable by specifying the conditions under which authorship may be said to obtain, and by situating AI-assisted textual production within those conditions.

This analytical distinction should be read alongside the earlier Arendtian framing of authorship. Linguistic labour and many components of linguistic work can be performed or supported by LLMs, but linguistic action remains tied to public responsibility, position-taking, and answerability. Authentic authorship therefore resides most clearly in linguistic action: appearing before others, assuming responsibility for a claim, and becoming answerable within a shared scholarly world.

The institutional perspective adds that authorship is not reducible to an inner property of the writer. It is authorised through procedures, norms, and gatekeeping practices that determine who may speak, how claims become recognisable, and how credit and responsibility are assigned. AI can reproduce many of the procedural markers of scholarly authorship, including genre conventions and disciplinary style, but this does not by itself amount to authorial attribution in the stronger sense of accountability.

The epistemic perspective clarifies why fluent textual output is insufficient for authentic authorship. Current LLMs can produce plausible and coherent discourse, but they do not assume epistemic responsibility for the claims they generate. Whether future systems might satisfy stronger conditions of understanding, grounding, or subjective position-taking remains open; the present argument concerns the limits of current LLMs as authorial agents.

Authentic authorship does not require perfect fulfilment of all these conditions. Rather, it presupposes the capacity to take responsibility for claims, to relate those claims to a shared world, to sustain commitments over time, and to remain answerable within institutional and disciplinary practices. In this sense, authenticity is not a property of textual fluency but of responsible authorial positioning.

These positions sit comfortably with the view of an authentic author as neither merely sincere, expressive, political, nor moral, but as one who adopts a rationally defensible position within a shared epistemic framework (Chakrabarty, 1998). Plausibility is not understood here as access to absolute truth, but as intelligibility within historically and institutionally produced criteria of evidence, argument, and reality. Authenticity is therefore not unlimited self--expression; authorial action must remain epistemically legible and publicly contestable.

In summary, authenticity in authorship is situated at the intersection of Arendtian public responsibility, Foucauldian discursive authority, and epistemic defensibility. Whereas AI can assist linguistic labour and linguistic work, conform to institutional genre markers of authorial voice, and produce plausible academic discourse, it cannot be an authentic author in the strong sense developed here because it cannot own a stance, sustain commitments over time, or be accountable as an author embedded in scholarly practices. Authenticity in AI--mediated scholarly practice is therefore best understood as responsible self--ownership of epistemic commitments, enacted as public answerability within disciplinary criteria of plausibility.

The proposed framework also shows the interrelationships among these propositions. If authenticity is reduced to romantic self-expression, almost any linguistic output might count. If institutional authority alone is privileged, only

dominant voices count. If epistemic relativism is accepted, no adjudication is possible. For adjudication to remain possible, plurality and public responsibility must coexist with shared criteria of reality. This is why authentic authorship requires the intersection of praxis, institutional recognition, and epistemic responsibility.

AI does not threaten public reason simply by undermining facticity; it threatens it by displacing human responsibility for sustaining and contesting epistemic standards. It can stabilise the form of epistemic reasonableness, such as evidence, argument, citation, and coherence, without securing the substance of epistemic validity. AI hallucination illustrates the same distinction: a system may reproduce the form of scholarly plausibility while lacking grounding and responsibility for error. The concern, then, is epistemic automation without corresponding responsibility.

The conditions proposed here are not fixed once and for all. As AI capabilities develop, debates about understanding, grounding, agency, and subjective experience may shift. The question of authenticity in authorship should therefore be treated as dynamic rather than finally resolvable at a single technological moment.

Conclusion and future research

This paper has been primarily conceptual, with the aim of developing a framework for approaching the question of whether AI can be considered an authentic author. The argument has distinguished authorship from mere textual production by locating authenticity at the intersection of praxis, institutional recognition, and epistemic responsibility. From the Arendtian perspective, AI can perform linguistic labour and support aspects of linguistic work, but cannot itself initiate linguistic action. From the institutional perspective, AI can reproduce many of the procedural and genre-based markers through which authorship is recognised, but it cannot itself assume attributional responsibility. From the epistemic perspective, current LLMs can generate plausible and coherent discourse, but they do not satisfy the stronger conditions

of grounding, epistemic commitment, and public answerability required for authentic authorship.

Future research could examine how the distinction between linguistic labour, linguistic work, and linguistic action operates in concrete scholarly settings. One direction would be to analyse institutional AI policies as boundary-making texts that distinguish author from assistant, permissible support from inadmissible substitution, and procedural compliance from authorial responsibility. A second direction would be to examine drafting, peer review, and revision processes to identify where AI performs linguistic labour, where human judgement organises linguistic work, and where authors assume epistemic responsibility through linguistic action. Such work could use version histories, tracked changes, peer review reports, supervisor feedback, or author reflections to examine how responsibility and position-taking are enacted in practice.

Further work could also examine how disciplinary, institutional, and cultural contexts shape expectations of authentic authorship. The present article has not offered a comprehensive sociocultural account of authorship and authenticity; rather, it has developed a conceptual vocabulary for distinguishing textual production from authentic authorship as embodied, institutionally recognised, and epistemically answerable action. Since AI capabilities and institutional norms will continue to change, the question of authenticity in authorship should be understood as dynamic rather than finally resolved at a single technological moment.

References

Arendt, H. (1958). *The Human Condition*. Chicago: University of Chicago Press.

Barthes, R. (1967). The death of the author. In: *Image, music, text* (pp. 142–148). London: Fontana.

Bender, E. M., McMillan-Major, A., Gebru, T., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? *FAccT*

'21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). DOI: 10.1145/3442188.3445922.

Burke, S. (1995). *Authorship: From Plato to the Postmodern (A reader)*. Edinburgh: Edingurgh University Press.

Chakrabarty, D. (1998). Minority histories, subaltern pasts. *Postcolonial Studies*, *1*(1), 15–29.

Chalmers, D. J. (1995). Facing Up To Problem of Consciousness. *Journal of Consciousness Studies*, *2*(3), 200–219.

Chalmers, D. J. (2023). Could a Large Language Model be Conscious? *arXiv:2303.07103*. DOI: 10.48550/arXiv.2303.07103.

Chartier, R. (2003). Foucault's Chiasmus: Authorship between Science and Literature in the Seventeenth and Eighteenth Centuries. In: M. Biagioli, & P. Galison (Eds.), *Scientific Authorship: Credit and Intellectual Property in Science* (pp. 13–31). New York: Routledge.

Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origin and Use*. New York: Praeger.

COPE Council (2019). *COPE discussion document: Authorship*. Committee on Publication Ethics.

Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices? *Journal of the American Society for Information Science and Technology*, *52*(7), 558–569. DOI: 10.1002/asi.1097.

Foucault, M. (1969). What is an author? In: *Language, counter-memory, practice. Selected Essays and Interviews* (pp. 113–138). Ithaca, New York: Cornell University Press.

Frankfurt, H. G. (2005). *On Bullshit*. Princeton: Princeton University Press.

French, R. M. (2000). The Turing Test: The First Fifty Years. *Trends in Cognitive Sciences*, 4(3), 115–121.

Goddu, M. K., Noë, A., & Thompson, E. (2024). LLMs don't know anything: reply to Yildirim and Paul. *Trends in Cognitive Sciences*, *28*(11), 963–964.

Guignon, C. (2004). *On Being Authentic: Thinking in Action*. Routledge: London and New York.

Hicks, M. T., Humphries, J., & Slate, J. (2024). ChatGPT is bullshit. *Ethics and Information Technology*, *26*(38). DOI: 10.1007/s10676-024-09775-5.

Hyland, K. (2009). *Academic Discourse: English In A Global Context (Continuum Discourse, 7)*. London: Continuum International Publishing Group.

Jones, C. R., & Bergen, B. K. (2024). People cannot distinguish GPT-4 from a human in a Turing test. *arXiv:2405.08007*, 1–23. DOI: 10.48550/arXiv.2405.08007.

Kang, E., & Shaw, M. (2024). tl;dr: Chill, y'all: AI Will Not Devour SE. *arXiv:2409.00764*. DOI: 10.48550/arXiv.2409.00764.

Millière, R., & Buckner, C. (2024). A Philosophical Introduction to Language Models–Part I: Continuity With Classic Debates. *arXiv:2401.03910*. DOI: 10.48550/arXiv.2401.03910.

Millière, R., & Buckner, C. (2025). A Philosophical Introduction to Language Models–Part II: The Way Forward. *arXiv:2405.03207*. DOI: 10.48550/arXiv.2405.03207.

Nagel, T. (1974). What Is It Like to Be a Bat? *The Philosophical Review*, *83*(4), 435–450. DOI: 10.2307/2183914.

Puşcaşu, I.-d. (2024). Creating with AI: On recent debates about authorship revisiting the influence of Barthes and Foucault. *Studia Ubb. Philosophia*, *69*, 49–60. DOI: 10.24193/subbphil.2024.sp.iss.03.

Saeidnia, S. A., & Lang, A. (2017). *An Analysis of Hannah Arendt's The Human Condition*. London: Macat International Ltd.

Sakana.ai (2026, 28 January). Retrieved from https://sakana.ai/ai-scientist/.

Searle, J. R. (1980). Minds, brains, and programs. *The Behavioural and Brain Sciences*, *3*(3), 417–457. DOI: 10.1017/S0140525 × 00005756.

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

Yildirim, I., & Paul, L. (2023). From task structures to world models: What do LLMs know? *arXiv:2310.04276v1*. DOI: 10.48550/arXiv.2310.04276.