Earned Trust: Verifiable Disclosure as a Credential-Substitute for AI-Assisted Work

Introducing the AIast disclosure, with an optional Lost Innovator (LI) prefix

William Stafford, ADN, LI-AIast5 · Independent Researcher

Version 1.8 — July 2026 *Preprint. Zenodo concept DOI: 10.5281/zenodo.20719927 (this DOI represents all versions of the standard).*

This version's changes, and the full record of earlier versions, are in the Version history section preceding the References.

Abstract

Institutional credentials function as *pre-paid trust*: a reader extends a degreed, affiliated author the benefit of the doubt because a recognized body has already vouched for them. People without those credentials have had no comparable way to earn a hearing, and their work has too often gone unread regardless of merit. This paper argues that generative AI changes that situation in two directions at once. It is widely told as a story of *replacement* — machines displacing knowledge workers — but it is also, less visibly, a story of *recruitment*: the same tools let people who could never have produced formal work begin to produce it. The question that follows is how such work can be trusted when its author carries no credential. The answer proposed here is **earned trust through verifiable disclosure**: where a credential buys trust in advance, transparency can earn it afterward, by making exactly how a work was produced open to inspection. The paper offers a concrete instrument for this — the **AIast disclosure**, a compact, controlled-vocabulary record of which AI services were used, what each did, and how the work was checked, backed by producible evidence — and an optional, field-relative **Lost Innovator (LI)** prefix for authors working outside the credentialing system. It situates the instrument against existing practice (the CRediT taxonomy; publisher AI-disclosure policies), explains why structured disclosure beats prose, gives the incentives for credentialed and uncredentialed authors alike, catalogs the ways the standard can be abused, and states its limits. The central claim is modest in form and large in consequence: trust need not always be borrowed from an institution; it can be earned by showing the work. The paper is itself disclosed under the standard it proposes.

Keywords: epistemic trust; credentialism; AI disclosure; research integrity; contributorship; transparency; CRediT.

1. The trust problem

When a reader trusts a paper they have not verified, they are usually trusting something other than the argument: the author's credential. A doctorate, a university affiliation, a professional license — each is a token issued by an institution that has, in effect, pre-paid the reader's trust. The reader need not check the work, because the gate already did. This is efficient, and mostly defensible. Credentials are neither infallible nor sufficient — gates fail, and a degree certifies training rather than correctness — but they remain the default mechanism by which a reader's trust is allocated. They are also a gate, and gates exclude.

Those without the token have had no equivalent route to a hearing. An idea from someone without the right degree, the right post, or any institutional affiliation has too often died unread — not necessarily because it was weak, but because its author had no way to purchase the trust that gets work taken seriously. Some of that excluded work was surely worthless. Some of it was not, and we have no way to count what was lost.

Generative AI is reshaping this in two directions, and the public conversation has noticed only one. The visible story is *replacement*: AI displacing the people who do knowledge work. The quieter story is *recruitment*: the same tools lowering the barrier so that people who could never before have produced a formal argument — for want of training, fluency, time, or access to the literature — now can. Both stories are true. This paper is about the second, and about the problem it creates. If new contributors arrive without credentials, on what basis can their work be trusted?

The answer proposed here inverts the logic of the credential. A credential buys trust *in advance*; verifiable transparency can earn it *afterward*. Instead of "an institution vouches for me, so you need not check," the uncredentialed author says "I will show you exactly how this was made, what tools I used, what each did, and what I checked — and you, or anyone, can audit it." Trust is no longer borrowed from a gatekeeper; it is earned by exposure. The claim is bounded, and the bound matters: transparency substitutes only for the *trust-allocating* function of a credential — the part that gives a reader grounds to engage rather than dismiss — not for everything a credential signals. A degree also certifies training, method, and exposure to criticism, none of which showing one's work can confer. What exposure supplies is the narrower thing a reader most immediately needs: a basis for trusting *this* piece of work, earned rather than borrowed.

What the disclosure offers, then, is not a diploma and makes no claim to expertise. It is a disclosure of *method* — of how the author's statements were arrived at — offered so the work can be judged and tested on its own terms. The trust it seeks is precisely bounded: trust enough to be *read and tested*, not trust that the author is right. This is how inquiry already works for the credentialed too. An author advances a claim; others test it; the claim is strengthened, qualified, or refuted. A claim later disproved has not failed the system — provided it was specific enough to be worth engaging, it fed the system by entering the record where others could take it up. AIast asks only for entry into that process and leaves the verdict to the process. The rest of this paper develops that claim and gives it an instrument.

2. The disclosure landscape

Two existing developments frame the instrument and explain the gap it fills.

CRediT. The Contributor Roles Taxonomy was introduced to move scholarly credit "beyond authorship" toward a granular account of who did what (Allen, Scott, Brand, Hlava, & Altman, 2014; Brand, Allen, Altman, Hlava, & Scott, 2015), and is now the standard ANSI/NISO Z39.104-2022 (NISO, 2022). It succeeded because it was short, used a controlled vocabulary, and was governed by a neutral body. But it describes *human* contributions and always will, because its roles presume an accountable agent. It has no vocabulary for a tool.

Publisher AI policy. Since 2023, major publishers and editorial bodies have converged on two rules: AI cannot be an author, because authorship requires accountability a tool cannot bear; and AI use must be disclosed. The ICMJE requires authors to disclose AI-assisted technologies in both the cover letter and the submitted work, locating writing/editing disclosures in the

acknowledgments and data/figure uses in the methods — guidance introduced in May 2023 and carried into its current January 2026 edition (ICMJE, 2026). Nature Portfolio's policy likewise denies LLMs authorship, requires documentation of their use in the methods, and exempts only routine copy-editing (Nature Portfolio, 2024).

The gap is the seam between them. CRediT standardizes human roles but not AI roles; publisher policy mandates AI disclosure but not its form; and neither makes the disclosure *checkable*. Meanwhile, undisclosed AI use in published work continues to be documented (Glynn, 2025). What is missing is a compact, comparable, evidence-backed way to say what the AI did — which is exactly what an author trying to earn trust without a credential most needs.

Since this specification was first drafted, the landscape has moved — toward classification, and toward the questions this instrument answers. Three developments now frame it.

STM classification. In September 2025 STM — the International Association of Scientific, Technical and Medical Publishers — finalized its *Recommendations for a Classification of AI Use in Academic Manuscript Preparation* (STM, 2025): a voluntary recommendations document — nine activity categories that publishers may use when setting policies and author-declaration requirements, spanning language refinement, manuscript drafting, translation, refinement of data reported in the manuscript, figures, visualizations of research data, reference gathering, presentation of code, and AI-generated content presented as research results. Its scope is manuscript preparation, not the broader research process, and it is a publisher-facing activity classification rather than a universal author mandate. It distinguishes forms of manuscript assistance, but it does not set out to provide a manuscript-level measure of overall AI involvement, a standard human-oversight statement, or an evidence-retention obligation — those sit outside its stated scope.

The AID Framework. Weaver's Artificial Intelligence Disclosure (AID) Framework (Weaver, 2024, 2026) approaches disclosure from the author's side, as this standard does — a voluntary author framework, a proposal like this one: a structured statement built on fourteen standardized headings — identifying the tools, versions, and dates of use, and organizing AI contributions across activities from conceptualization and data analysis to editing, translation, and project administration — with free-text descriptions beneath the relevant headings, in a format designed to be both human- and machine-readable. It has been proposed and discussed in library and educational settings, the bottom-up path any voluntary instrument realistically has. Its distinction from this standard is narrow and specific: as published in its 2024 introduction, it does not require the author to retain or produce underlying records — transcripts, logs, or source-audit documentation — to substantiate the disclosure.

The Global Reporting Standard process. The International Science Council, COPE, STM, and the Global Young Academy — with the World Conferences on Research Integrity Foundation supporting the process's Vancouver focus track — are developing a *Global Reporting Standard for AI Disclosure in Research*: an open consultation, with no adopted output yet. Its second consultation round, July through October 2026, addresses disclosure thresholds, placement and structure (including a research-task taxonomy), negative or null declarations, and accountability — verification, oversight, and audit trails. A third round, at the end of 2026 into early 2027, will refine the proposed standard against consultation feedback (International Science Council, 2026).

The wave is wider than these three. GAIDeT, a delegation taxonomy in the CRediT mold, has the author declare which research tasks were delegated to which named tool, rendered as a prose

declaration (Suchikova, Tsybuliak, Teixeira da Silva, & Nazarovets, 2026). Resnik and Hosseini give the threshold question its sharpest published formulation: disclosure should be mandatory only where AI use is intentional and substantial (Resnik & Hosseini, 2026). AMEE Guide No. 192 carries the question into health professions education, asking authors for a candid prose description of how generative AI shaped the work (Cleland, Driessen, Masters, Lingard, & Maggio, 2026). And outside the peer-reviewed literature, the AIR framework — a practitioner framework published through Figshare — grades AI involvement per research stage on a five-band scale (Young, 2026). They classify, threshold, and describe; none obliges the author to retain and produce the record behind the disclosure. This comparison is targeted rather than systematic — prominent policies and frameworks located through publisher, standards-body, repository, and peer-reviewed sources, current through July 2026 — and claims about comparators refer to this reviewed set.

Where this instrument sits. The overlap among these initiatives is substantial, and the honest claim is narrow. None of these — including this one — is an adopted standard: STM's is a voluntary recommendation, AID a voluntary author framework, the Global process an open consultation, and this a publicly deposited proposal. Two pieces of the present instrument are common property: sole human responsibility is standard across publisher policy, and disclosure thresholds are under active development in the Global consultation. Two are rare: the compact author-facing function notation, and the obligation to retain and produce the evidence behind the disclosure — neither of which the comparators, as published, carry. The proposal is the combination: notation, threshold, responsibility, and evidence obligation in one compact instrument. The approaches compose rather than compete: an activity classification can sit inside the workflow statement, and the evidence commitment can stand behind any of them.

3. The AIast disclosure

The instrument attaches to named human authors. Authors of record are always human; the instrument never implies an AI authored the work. It has two parts: a compact byline mark and a short workflow statement. Throughout, "the standard" names the instrument being proposed — shorthand, not a claim of adopted status (Section 9).

3.1 The mark

```
Byline: Name, [credentials,] [LI-]AIast(n)

AI-use disclosure: AIast — SERVICE(fn) · SERVICE(fn) · … + workflow
statement
```

The byline mark is compact so it can sit after a name like any post-nominal, beside conventional credentials: `J. Smith, BSN, LI-AIast5`, or `J. Smith, PhD, AIast2`. Here n is the number of distinct disclosed AI entries: each user-facing AI service, plus each self-hosted model used above the threshold (see the rule below the table). The full breakdown — which services, and what each did — lives with the workflow statement, placed where the venue requires — methods, acknowledgments, or a dedicated AI-use disclosure section with the references when nothing is prescribed. The LI prefix is optional (Section 5).

Table 1. Canonical service codes (extensible via the registry of Section 9)

Code	Service	Provider
CL	Claude	Anthropic
GP	ChatGPT	OpenAI
GR	Grok	xAI
GM	Gemini	Google
PX	Perplexity	Perplexity
CP	Copilot	Microsoft
MS	Mistral	Mistral
DS	DeepSeek	DeepSeek

Open-weight models with no first-party service — Llama, for example — are disclosed by the platform that served them, with the model named in the workflow statement; a model run on the author's own hardware still counts toward n and enters the breakdown as a self-hosted entry — Self-hosted Llama (d/c), for example — with the model and version or checkpoint reported as available to the author, since no service operator exists to expose or attest a version. Any service not in this list is written out in full.

Writing the breakdown. An entry names its service either in full — *Claude (i/d/c)* — or by its code — *CL (i/d/c)* — and both are conforming: the full name is the display form, the code is the compact form, and a single disclosure uses one form throughout. Tags sit in parentheses, lowercase, separated by slashes, in the order of the tag table; entries are separated by middle dots; the list opens with *AIast* —. A self-hosted entry takes the same form, prefixed Self-hosted. The worked examples in Section 6 and the disclosure on this paper use the display form.

Table 2. Function tags — the type of assistance, which is the load-bearing signal:

Tag	Role
i	ideation / interlocutor (dialectic, thinking-through)
d	drafting / composition
c	critique / adversarial review
v	verification / fact- and logic-checking
s	source retrieval

`PX (s)` discloses that AI was used *only* to find sources — light assistance. `CL (i/d/c/v)` discloses that AI was in the ideas, the draft, the critique, and the checking — deep assistance.

That distinction is what a reader actually needs. The tags classify the AI's role, not the task type: translation, code, or figure generation performed by AI is drafting (d) of that artifact, named in the workflow statement — where an activity classification such as STM's (Section 2) may be cited for task-level detail.

On the number. `n` is deliberately a *navigational flag*, not a metric. It tells a reader, at a glance, that AI was used and how many disclosed entries to expect in the breakdown — no more. It is explicitly **not** a measure of rigor, effort, or quality: `AIast5` is not "better" or "more careful" than `AIast2`, and one deeply used service may involve far more scrutiny than five used in passing. The substance lives in the function tags, the workflow statement, and the producible evidence; the count is only a pointer to them. Stating this plainly is part of the standard, because the count's one real failure mode is being read as a score. The number is accordingly **optional**: the function-tag breakdown in the references is the actual disclosure, and an author who finds the byline count unhelpful may omit it and write AIast alone — a byline choice only, since the breakdown must still appear with the disclosure. The count is retained only as a convenience for readers who want an at-a-glance signal — not because it carries information the breakdown lacks. The count refers to disclosed entries used in disclosure-worthy functions — each user-facing AI service, plus each self-hosted model above the threshold — not underlying models, model versions, or threshold-exempt software features. One service exposing several selectable models counts once, with the models noted in the workflow statement; the same model reached through two services counts twice, because the service is what the author used.

When the mark applies. The disclosure is for *substantive* assistance — AI contribution that shapes or replaces the author's judgment or materially influences the reported content: the work's ideas, argument, text, evidence, interpretation, figures, or code. Routine technical assistance that conventional tools have long performed — spelling and grammar correction, formatting, reference styling — falls below the mark only when it does not introduce or materially alter meaning, claims, reasoning, evidence, or structure — the boundary publisher policies draw for routine copy-editing; Nature Portfolio's, for example, exempts it only where there is no generative editorial work or autonomous content creation and human accountability for the final text remains (Nature Portfolio, 2024). Embedded or built-in AI features are judged by the same content test, and output consulted but discarded still counts when it shaped the author's thinking — that is what the i tag records. Verification does not retroactively exempt a use from disclosure; it is itself a taggable function. The practical test: would a disciplinary peer reasonably need to know about the AI use to understand how the work was produced or to assess its reliability? Cumulative effect counts — many small uses that together alter the argument meet the threshold even when no single one does — and when reasonable doubt remains, disclosure is preferred. In multi-author works the disclosure covers AI use by all authors and disclosed contributors; the submitting author is responsible for collecting it. An author who wishes to state the negative affirmatively may write *AIast0*: a scoped declaration that, to the author's or authors' reasonable knowledge, no AI use meeting this threshold contributed to the work. It does not assert that no embedded or threshold-exempt AI functionality existed anywhere in the tools used. Silence remains permitted, but silence is not a conforming disclosure and asserts nothing. A bare *AIast* with no breakdown is likewise not conforming: the mark never substitutes for the breakdown, it points to it. *AIast0* is the only form that stands without a breakdown, because there is nothing to point to.

3.2 The workflow statement

The mark is paired with a brief prose statement recording the *process* and pointing to the evidence: what each service did, the depth of iteration, what was checked and by whom, and that transcripts are available. For example:

Method: drafted in collaboration with Claude (model version, date) over several rounds; argument separately critiqued by GPT and Gemini; every factual claim source-checked via Perplexity and against primary sources. The author directed the work, integrated all contributions, and is solely responsible for every claim. Working transcripts available on request.

The closing clause — *available on request* — is the operational heart of the standard. Everything the disclosure asserts is only as good as the evidence the author can produce. One word here is chosen deliberately: critique by another AI service is separate, not independent — different services can share training material and blind spots — and this standard reserves independent for review by an uninvolved human.

3.3 The evidence package

“Available on request” is a commitment, and this section states what it commits to. The retained record behind a disclosed work has four parts: the working transcripts or logs from each disclosed service; a session manifest — the first page of the evidence — listing each working session by date, service, and role, recording the model version exactly as the service exposed it and writing *not exposed* where it did not, since a manifest that guesses is worth less than one that says plainly where the record ends; the critiques received and the author’s rulings on them, including decisions to decline specific recommendations; and the source-verification records for the claims the author checked.

The record is kept for at least five years from publication — longer where a legal, institutional, funder, or venue requirement applies. Production may be scoped to the challenge: a reader doubting one claim is answered with the sessions that bear on that claim, not handed the archive. What is produced must be produced honestly — an excerpt is a window, not a curation.

Two things may be withheld: material held under a duty of confidence to others — including duties imposed by law, contract, or research-ethics requirements — and personal data. Redaction is the exception, not the norm, and the judgment is the author’s: before withholding anything, the author weighs the need to withhold against what the withheld material contributed to the work. The more a withheld passage shaped the work, the more the redaction costs — a reader who cannot inspect what mattered may rightly discount the claims that rested on it, and if peer review finds that a redaction collapses the argument, that is the price of the redaction. Each redaction is marked where it occurs and states its reason; a marked redaction is part of the record, not a hole in it. Nothing in this standard alters the ordinary rules of authorship, citation, and permission that govern all scholarly work.

An author who wants the record to be tamper-evident may seal it: publish, with the work itself, a cryptographic fingerprint — a hash — of each retained file. A file produced years later that matches its published fingerprint is byte-for-byte the file that existed on publication day — not reconstructed, not trimmed. The manifest names the algorithm (SHA-256, for example), each file, its size, and its digest. Sealing has a precise limit: it proves a record unchanged since publication,

never that the record was complete when sealed. It answers the altered-record abuse in Section 8; the omission asymmetry stands.

3.4 Conformance

This section states, in one place, what a conforming disclosure must contain and what happens when a requirement is not met. Elsewhere in this paper, requirements are stated in ordinary prose; this section restates the load-bearing ones using a fixed vocabulary. MUST and MUST NOT identify a requirement whose absence makes a disclosure non-conforming. SHOULD and SHOULD NOT identify a strong recommendation that a disclosure may depart from for a stated reason without becoming non-conforming. MAY identifies something left to the author's discretion. Nothing in this section overrides Sections 3.1–3.3; it summarizes them for reference and resolves cases the prose leaves ambiguous.

A conforming positive disclosure MUST include either the byline mark in one of its authorized forms (Section 3.1) or, where venue rules prohibit nonstandard byline additions, a standalone manuscript-level AIast mark placed with the disclosure; a breakdown listing every disclosed service and its function tags, using either the full name or the code consistently within that disclosure, MUST NOT mix the two forms in one breakdown; and a workflow statement addressing what each service did and that the record is available on request (Section 3.2). A disclosure SHOULD state the depth of iteration and what was checked and by whom, and SHOULD name the model or model version where available; it MAY omit either when the interaction genuinely had none to report — a single search pass, for example — provided the statement says so rather than remaining silent on the point. The workflow statement MUST identify the named human author or authors as having directed and integrated the work and as retaining responsibility for its claims; the wording MAY vary, but human responsibility MUST be explicit.

AIast0 (Section 3.1) is the one conforming disclosure with no breakdown: it MUST be a scoped negative declaration — to the author's or authors' reasonable knowledge, no AI use meeting the substantive-assistance threshold contributed to the work — and it MUST NOT be used merely because a breakdown would be inconvenient to compile. A bare AIast or LI-AIast with no breakdown and no AIast0 declaration is not conforming under any circumstance; the mark always points to a breakdown or to the AIast0 declaration, never to nothing.

On the count. n MUST equal the number of distinct entries disclosed in the breakdown — user-facing services and self-hosted models alike (Section 3.1) — counting the same underlying model reached through two different services as two entries, because the count tracks what the author used, not models. An author MAY omit the count and write the bare byline mark AIast, provided the breakdown still appears in the disclosure; omitting the count MUST NOT be read as omitting the breakdown. A service absent from the canonical table (Section 3.1) MUST be written out in full rather than assigned an unregistered code; once the registry of Section 9 exists, an author MAY apply for a permanent code for a service used repeatedly.

Multi-author works. The disclosure is manuscript-level by default: it covers all threshold-meeting AI use by every author, and the submitting author collects and holds the evidence package (Section 3.1). A byline mark attached to an individual author counts that author's disclosed entries; a standalone manuscript-level mark counts all disclosed entries for the work. Author-specific breakdowns MAY be added where use differs materially.

On evidence. The retained record MUST include the four evidence categories of Section 3.3 to the extent each applies to the work: for every disclosed entry, the relevant transcripts or logs and manifest entries; critiques and rulings where critique occurred; and source-verification records where sources were retrieved or checked. The applicable record MUST be kept for at least five years from publication, or longer where a stricter retention requirement applies. Production in response to a challenge MAY be scoped to the claim under challenge rather than the full archive, but what is produced MUST be produced honestly and MUST NOT be curated to omit unfavorable material within the scope requested. Redaction is permitted only for the two categories in Section 3.3 — material held under a duty of confidence to others, and personal data — and each redaction MUST be marked where it occurs and MUST state its reason; an unmarked redaction is a failure mode under Section 8, not a conforming use of the exception. Sealing the record by published fingerprint (Section 3.3) is optional and MAY be adopted by any author; it is never required for conformance.

When evidence cannot be produced. If a reader or editor makes a specific, good-faith request for material within the scope described above and the author cannot produce it, the disclosure it supported becomes non-conforming and unsubstantiated for the claims the missing evidence was meant to back — not automatically false, since the underlying claim may still be true, and not automatically fraudulent, since records can be lost as well as fabricated. What a reader is entitled to conclude is that the specific claim no longer carries the evidentiary backing this standard exists to provide, and may weigh it accordingly.

Two minimal specimens illustrate the floor. Conforming: “Byline: A. Author, AIast1. AI-use disclosure: AIast — CL (c). Method: the draft was separately critiqued by Claude in a single round; all writing and revision by the author, who is responsible for the work. Transcript retained and available on request.” Non-conforming: “Byline: A. Author, AIast3. AI-use disclosure: AIast.” — the count is present but the breakdown is bare, which Section 3.1 excludes for any mark other than AIast0.

4. Why structured disclosure beats a prose paragraph

Why not simply write a sentence? Three reasons. **Comparability:** structured fields can be read across many works at once and compared at a glance; prose buries the distinction in varying phrasing. **Completeness:** a controlled vocabulary prompts the author to address each role, where prose lets convenient omissions pass. **Checkability:** the structured mark commits to specific, falsifiable claims — *this service did this* — each backed by retained evidence, whereas “AI helped with editing” commits to almost nothing. Prose remains fine for a single human reader; the case for structure is the case for a shared convention that scales, compares, and can be audited.

5. The optional LI prefix

LI (Lost Innovator) is an **optional, self-applied, descriptive** prefix. A Lost Innovator is *an author working without the credentials or institutional standing that the relevant field conventionally treats as a license to author or conduct research independently — whether they hold no degree, a degree below that threshold, a degree in a different field, or no academic affiliation.* The status is **relative to the work at hand**: a doctorate in physics does not, by itself, establish expertise in theology, and an associate’s degree in nursing is a genuine clinical qualification that does not, by itself, establish advanced research training in nursing scholarship,

and does not transfer to a theory of history. The designation identifies that credential-to-subject relationship; it does not declare the author unqualified to reason, research, or contribute. The prefix is not a qualification and makes no claim about the worth of the work. It explains *why* the author is leaning on verifiable transparency — because the credential that normally buys a hearing in this field was never on offer. One conventional route past the gate is collaboration with a credentialed co-author who shares responsibility for the work; the LI-AIast disclosure offers a route for the author who remains solely responsible: earning a hearing by showing the work rather than borrowing a credential.

Three properties keep the prefix from becoming a contested status symbol. It is **optional** — the disclosure is complete without it, and most users may never apply it. It is **defined** — there is a published definition to self-assess against, so "who decides?" has an answer: you do, against the definition, and the work backs you up. And it is **descriptive, not evaluative** — it claims nothing about quality, which the transparency and the work carry entirely. Because validity never rests on the label, the label cannot be inflated into a credential. One honest limit: unlike the AI-use claims, which are evidence-backed, the LI self-assessment is subjective and cannot be independently audited; it is offered as description, not certifiable status.

6. Worked examples

Example A — a deeply assisted paper by an uncredentialed author using four services:

Byline: **J. Author, LI-AIast4** AI-use disclosure: AIast — Claude (i/d/c) · *ChatGPT (c) · Gemini (c) · Perplexity (s). Method: argument developed in dialogue with Claude and drafted with it over multiple rounds, Claude also serving as internal critic; separately critiqued by ChatGPT and Gemini; sources checked via Perplexity. The author directed the work and answers for its claims. Transcripts available on request.*

Example B — a lightly assisted paper by a credentialed author:

Byline: **A. Professor, PhD, AIast1** AI-use disclosure: AIast — Perplexity (s). *Method: AI used only for literature search, in a single pass; every source it returned was read and checked by the author; all writing and analysis by the author, who directed the work and answers for its claims. Search records retained and available on request.*

Example C — a verification-heavy case using three services:

Byline: **C. Researcher, AIast3** AI-use disclosure: AIast — ChatGPT (d) · *Claude (c/v) · Perplexity (s/v). Method: drafted with ChatGPT over successive rounds; separately critiqued and fact-checked by Claude; sources retrieved and checked with Perplexity. The author integrated the outputs and remains responsible for the work. Transcripts and source-audit records retained and available on request.*

Example B shows that the mark is universal — a credentialed academic discloses AI use beside the real degree, without adopting an identity that does not fit. Example A shows the core use and the number working as designed: `AIast4` flags "four services, see the breakdown," and says nothing about rigor.

Example D shows the standard under pressure — a *challenged* disclosure. Suppose a reader doubts Example A: the disclosure says Claude was in the ideation, the drafting and the critique —

but did that ideation in fact amount to supplying the core argument the author now claims as their own? The tags record the type of a contribution, not its depth, so the challenge tests the disclosure as a whole. Under the standard the challenge has a procedure rather than a shouting match. The author is asked to produce the working transcripts, and the record shows whether the disclosure — the `(i/d/c)` tags with their workflow statement — fairly represented the contribution or understated it. The disclosure converts an unfalsifiable suspicion — "AI probably wrote this" — into a checkable question with an answer. That is the enforcement model in miniature: not prevention, but *exposure on demand*.

7. Why it earns trust — and for whom

The thesis of Section 1 is that transparency can earn the trust a credential would otherwise buy. This section makes that concrete and meets the obvious objection.

For the uncredentialed (the Lost Innovator). This is the primary case. Lacking a token to buy trust in advance, the author offers the alternative: bounded, evidence-backed legibility. "I cannot show you a diploma, but I can show you precisely how this was made, what each tool contributed, and what I checked — and here are the transcripts." That is a different kind of warrant — an earned one — and it is available to anyone willing to work in the open.

For the credentialed (everyone else). The incentive is not only for outsiders, which is what makes adoption plausible. A credentialed author has three distinct reasons to disclose well. First, compliance: many journals now require AI-use disclosure (ICMJE, 2026; Nature Portfolio, 2024), and a structured mark meets that requirement more cleanly than ad hoc prose — subject to each journal's own placement rules, cover-letter requirement, and submission forms. Second, **reputation**: in a climate of active suspicion about hidden AI use, proactive, checkable disclosure is insurance — the author with nothing to hide says so in a form that can be verified. Third, **risk management**: for clinicians, lawyers, and consultants, a structured record of what AI did and what the author verified is due-diligence documentation against the day a claim is contested. The standard's logic generalizes: in a climate of suspicion, the author who discloses well leaves nothing for readers to assume.

A limit worth stating plainly: transparency is not expertise. Showing one's work cannot manufacture the domain knowledge, training, and methodological fluency a credential also signals; an honest, fully disclosed argument can still be the work of someone who does not understand the field. The claim is therefore narrow. Transparency substitutes for the credential's role in *allocating trust* — in giving a reader grounds to engage rather than dismiss — not for the competence the credential also certifies. It buys a hearing, not a verdict. What an author does with that hearing still depends on whether the work is any good — and "any good" is not the author's to certify; it is settled downstream, by whether the work survives the scrutiny of others. A disclosed claim that is refuted has still done its job: it entered the testable record and provoked the response that refuted it. AIast is therefore a *complement* to the ordinary means of earning a competence-hearing, never a substitute for expertise. It gets the work into the room; what happens there is decided by the work and its critics, exactly as for any other author.

A second limit matters just as much: earned trust is not the same as a warm reception. Recent experiments find that disclosing AI use can *lower* a reader's immediate trust rather than raise it — thirteen experiments, across tasks and audiences, with the effect running through perceived legitimacy (Schilke & Reimann, 2025). The standard does not dispute this, because its claim was

never that disclosure makes an author better liked. The distinction is between *perceived* trust — what a reader feels on contact; *warranted* trust — what a reader has inspectable grounds for; and *correctness*, which only scrutiny settles. AIast is built for the middle term. It cannot promise a favorable reception, and a disclosed work may at first be received more coldly than a silent one. What it offers is the warrant: the grounds to engage rather than dismiss — a hearing, not a verdict. And the same experiments supply this section's insurance argument with its evidence: in those experiments, exposure as an undisclosed AI user produced a larger trust penalty than voluntary disclosure did (Schilke & Reimann, 2025). Between the modest immediate cost of candor and the larger deferred cost of exposure, the standard's bet is the cheaper one.

The objection: does transparency work if no one checks? A fair challenge to the whole thesis is that most readers will never request the transcripts, so the disclosure is theater. The reply has three parts. First, checkability disciplines the *author* regardless of whether anyone checks: people tend to write more carefully when they know they can be audited, exactly as open data and open-source code improve quality through exposure even though few readers inspect them. Second, the credible possibility of verification can deter fabrication; a bluff that can be called is a weaker temptation than one that cannot. Third, trust shifts from *who vouched* to *what was exposed*, which means a single motivated skeptic — a reviewer, a rival, a journalist — can act on the evidence even when the casual reader does not, and the knowledge that they might is often enough to change behavior. Transparency does not require universal checking to work; it requires that checking be *possible* and occasionally *actual*. That is a far lower bar than universal verification, and it is the same bar open-science practice has largely cleared.

Three practical objections, briefly. *Doesn't this burden authors and editors?* The burden is real but bounded: a compact public disclosure backed by a structured private record. The manifest, the rulings, and the retention floor are work beyond the prose acknowledgments publishers already request — and still far cheaper than being quietly disbelieved. *Must transcripts be made public?* No. "Available on request" means *producible*, not *posted*: an author may share relevant excerpts, redact unrelated material, or supply the record privately to an editor or challenger. The standard asks for checkability, not for exposing everything ever typed. *What about proprietary, internal, or fine-tuned tools that cannot be named precisely?* Disclose by category and provider where the exact system cannot be identified — for example, "internal fine-tuned model (provider)" — and write it out in full; once the registry of Section 9 exists, such services can be assigned codes. The principle is to name what can be named and be honest about what cannot, not to demand impossible precision.

8. Failure modes: how the standard can be abused

A disclosure standard that does not confront its own abuses is naïve, and a reader is already imagining them. The principal ones:

- **Underreporting.** An author lists a service for "editing" when it in fact drafted the argument. The function tags can lie like any self-report.
- **Exaggerated verification.** An author claims sources were checked when they were skimmed, or claims independent critique that was a single approving pass.
- **Selective disclosure.** An author lists the flattering services and omits the one that did the heavy lifting, or omits a service entirely.

- **Phantom transcripts.** An author writes “available on request” while having no usable record, betting that no one will ask.
- **Altered or incomplete records.** An author edits, selectively exports, or reconstructs the interaction history and presents what remains as complete. Sealing (Section 3.3) makes alteration after publication detectable; omission before it is the asymmetry discussed below.
- **Transparency theater.** An author piles up services and tags to *look* rigorous, exploiting the very misreading of the count that Section 3.1 warns against.
- **Motivated redirection.** An author steers the AI to confirm a predetermined conclusion rather than to test it, and discloses only the confirming exchanges. The author rightly controls the *direction* of the inquiry — the hypothesis is theirs, and the tool should not be left to wander into a different paper — but the disclosure’s worth depends on the tools having been used to *test* that hypothesis, including to attack it, not merely to manufacture support. Suppressing the disconfirming passages falsifies the record as surely as inventing a citation does. A transcript in which the AI only ever agrees is a warning sign, not a credential.

None of these is fully preventable, because all disclosure rests ultimately on honesty. But the standard narrows what can be *hidden*: every claim made points to producible evidence, so an unsupported claim can be exposed when the evidence is requested and cannot be produced or does not substantiate it. Omission is the harder case — evidence cannot reveal what was never disclosed — and that asymmetry is a real limit rather than a solved problem. This is the same enforcement model as open data — most submissions are not audited, but the standing possibility of audit, plus the reputational cost of being caught, does the work. The standard’s honest position is therefore not “this cannot be gamed” but “this makes every claim checkable and puts the burden of evidence on the discloser.” A norm of occasional spot-checking — by reviewers, editors, or readers — is what keeps it honest, exactly as it does for the rest of the scholarly record.

9. Governance

Two points are load-bearing; the rest can wait for adoption that has not yet occurred. **Open license:** this specification is CC BY 4.0; anyone may use, adapt, or extend it with attribution. Restricting it by copyright or trademark would defeat its purpose, since a format others cannot freely use cannot become shared. **A neutral code registry: service codes must mean** the same thing on every author’s work, so the canonical list should eventually live in a single public registry, new services added as two-letter codes, following CRediT’s path from community proposal to NISO standard. Until then, Section 3’s table is the provisional code list, version 1.8.

Admission to that registry follows a single test: *who runs the inference?* A service is the user-facing product through which an author reaches a model — the party that serves the computation and answers for the interface. A provider, a model family, or an open-weight checkpoint is not a service until someone serves it. New services enter by request, from an author who used one or from the service’s own operator, with evidence that the product is real, user-facing, and distinct from entries already listed. Codes, once assigned, are permanent: a service that shuts down or is renamed keeps its code, so that no published disclosure ever changes meaning. Listing is registration, not endorsement — a code asserts that a service exists and is distinctly nameable, and nothing else. And the registry can never block an author: an unlisted service written out in full is always conforming, so a disclosure is never held hostage to the registry’s pace.

Apart from its open license and registry specification, this remains a publicly deposited *proposal* rather than an adopted standard: as of July 2026 it has a citable, versioned deposit (Zenodo, DOI above), no documented independent adopters, and no public community-governance repository. The present version has been prepared for submission to the *Global Reporting Standard* consultation described in Section 2; whatever that process ultimately adopts, this instrument remains openly licensed and independently usable. The intended next step is a public repository for registry entries, code proposals, version history, and community contributions. A standard becomes real through use and contribution, not through a document declaring it so.

10. Limitations

This is a *proposal*, not a validated or adopted standard; its value is contingent on uptake that has not happened. The function tags are a simplification and may prove too coarse or too fine; revision through use is expected. The disclosure describes *process, not merit*: a fully and honestly disclosed work may still be wrong, and an undisclosed work may be excellent — disclosure is a basis for warranted engagement, not a measure of merit and no promise of perceived trust. The trust-substitute thesis assumes a culture in which verification is at least occasionally exercised; where no one ever checks and no reputational cost attaches to being caught, the mechanism weakens toward theater, as Section 7 concedes. And the LI prefix, being self-applied and broadly defined, will sometimes be claimed by those whose work does not merit attention — but since it carries no validity of its own, such misuse does not validate the work — the producible evidence package and the work itself remain the only things that matter — though it can still cost readers and editors attention and review time. The LI designation has, to the author's knowledge, exactly one applied case — its author — and has not been tested against a contested or borderline claim.

11. Conclusion

The credential is a pre-paid trust token, efficient for those who hold it and a closed door for those who do not. Generative AI is quietly recruiting a new class of contributors who arrive at that door without the token, and the question of how to trust their work is now live. This paper's answer is that trust need not always be borrowed from an institution; it can be earned by showing the work — and that AI-assisted writing, of all things, is where this becomes both necessary and feasible, because the tools that lower the barrier also create the disclosable record. The question behind it — what happens when AI lets people outside institutions take part in formal knowledge creation — is the one pursued elsewhere as the *lost innovators* hypothesis; this disclosure is its practical mechanism, the way a contributor arriving without a credential can ask to be read. The AIast disclosure is a minimal, openly licensed instrument for that earned trust: it makes legible what AI contributed, commits the claims to producible evidence, and lets an uncredentialed author offer transparency where a credential is not available — while serving any author who wants to give a clear, specific, checkable account of AI assistance. It is offered, not imposed. Whether it earns adoption will be decided by whoever finds it useful enough to apply and honest enough to defend.

Glossary and notation key

- **AIast** — the disclosure that a work was produced with AI assistance, by a human who directed it and remains answerable for it as an author of record. An AI is never an author of record. Pronounced like *essayist*.

- **AIast(n) — the byline form; n is a metavariable for the number of distinct disclosed AI entries — user-facing services and self-hosted models alike — and a rendered mark concatenates the count without parentheses: AIast5, not AIast(5).** A neutral, checkable navigational flag — **not** a measure of rigor, effort, or quality. An author may write *AIast0* — a scoped declaration that no threshold-meeting AI use contributed to the work (Section 3.1).
- **LI (Lost Innovator)** — optional, self-applied prefix: an author working without the credentials or institutional standing the relevant field treats as a license to author or research independently (no degree, a sub-threshold degree, an out-of-field degree, or no affiliation). Relative to the work at hand; descriptive, not a qualification.
- **Earned trust** — trust extended on the basis of *what an author has exposed to inspection*, as opposed to *pre-paid trust* extended on the basis of an institutional credential.
- **Verifiable method disclosure** — a record of how a work was produced that the author can substantiate on request (e.g., via transcripts); the disclosure makes the claim, the evidence makes it checkable.
- **Service codes** — `CL` Claude · `GP` ChatGPT · `GR` Grok · `GM` Gemini · `PX` Perplexity · `CP` Copilot · `MS` Mistral · `DS` DeepSeek.
- **Function tags** — `i` ideation/interlocutor · `d` drafting · `c` critique/adversarial review · `v` verification/fact- and logic-checking · `s` source retrieval.

Disclosure (per this standard)

Byline: **William Stafford, ADN, LI-AIast5**

AI-use disclosure (placed with the references): *AIast — Claude (i/d/c/v/s) · ChatGPT (d/c) · Gemini (c) · Grok (c) · Perplexity (c/s)* — tags: i ideation/interlocutor · d drafting · c critique · v verification · s source retrieval. This paper was produced with Claude (Anthropic, successive models 2025–2026) as interlocutor, drafting assistant, critic, verification assistant, and search agent, and was separately critiqued across drafts by ChatGPT (OpenAI), Gemini (Google) and Grok (xAI), whose adversarial reviews materially shaped it; specific model versions were not consistently exposed by the services and are recorded in the retained transcripts where available. For version 1.6.3, Perplexity (Perplexity AI, July 2026) compiled a comparison of eight organizations' AI-disclosure requirements that informed Section 2, and the author verified each claim used against the cited primary sources; the comparison and its correction history are retained with the working records and published on the author's website as a dated snapshot. The author directed the work, selected and rejected AI outputs, made all editorial decisions, and is solely responsible for the accuracy, interpretation, and conclusions of the work. For version 1.7, successive separate adversarial reviews of deposited version 1.6.3 and of the pre-deposit drafts were run with Perplexity, ChatGPT, Gemini and Grok. Perplexity identified a production defect in Section 10 that the author's own verification had missed; the remaining review rounds produced the further corrections listed in the change note. Every finding was checked against the primary sources before acceptance, and findings that proved incorrect were rejected on the record; the rulings are retained with the working records. For version 1.8, ChatGPT drafted the initial text of the conformance section (Section 3.4) in a review exchange directed by the author, who adopted it with revisions — its tags add d accordingly — and successive separate adversarial review rounds with ChatGPT shaped the remaining corrections; accepted findings were verified before adoption, and declined findings are recorded with their reasons. Working transcripts and version records are retained and available on reasonable request. The count of five reflects the user-facing AI services used in disclosure-worthy functions — a count of services, not a measure

of rigor. The count rose from four at version 1.6.3 with the addition of the fifth reviewing service named above; the increase records additional use, not additional merit.

License

Version history

Entries preserve the wording of the version they record.

Version 1.8 — changes from version 1.7. Additions: a notation rule in Section 3.1 (service codes and full names both conforming; one form per disclosure); a counting rule for self-hosted models, which now increment n and enter the breakdown as self-hosted entries; the registry admission test and process in Section 9 (who runs the inference; permanent codes; registration, not endorsement; never blocking); a new Section 3.3 defining the evidence package — the four-part retained record, a five-year retention floor, production scoped to the challenge, redaction narrowed to two marked categories and stated as the exception rather than the norm, and optional sealing of the record by published cryptographic fingerprint; a corresponding altered-records failure mode in Section 8; a distinction in Section 7 between perceived trust, warranted trust, and correctness, engaging the experimental finding that disclosure can lower immediate perceived trust (Schilke & Reimann, 2025), whose own results supply the insurance argument; and four further instruments acknowledged in Section 2 (GAIDeT; Resnik & Hosseini; AMEE Guide No. 192; the AIR framework), none of which carries an evidence obligation. Corrections: several empirical claims are qualified; one coercive sentence in Section 7 is softened; function-tag labels are aligned across the table, the Glossary, and the disclosure; residual British spellings are standardized; the byline-level and disclosure-level uses of the bare mark are distinguished; the registry-code instruction is deferred until the registry exists; AIast(n) is defined as metasyntax; the change note wording about the version 1.6.3 defect is made precise; one reference title is corrected to the publisher's form (Information Services and Use); and the version history moves to this section, leaving the abstract first. The worked examples now state iteration depth, what was checked and by whom, and the author's responsibility. A conformance subsection (Section 3.4) states the load-bearing requirements in a fixed normative vocabulary, resolves the bare-mark and evidence-failure cases, adds a manuscript-level multi-author rule, and closes with minimal conforming and non-conforming specimens. From two further review rounds: the count is restated over disclosed entries — user-facing services and self-hosted models alike; critique by another AI service is described as separate rather than independent, with independent reserved for human review; the withholding vocabulary is distinguished from cryptographic sealing, the confidence category is bridged to legal, contractual, and research-ethics duties, and the fingerprint claim is stated precisely; the author burden is restated honestly; the disclosure label is made venue-neutral; total legibility became bounded, evidence-backed legibility; the tag illustrations are written in the defined notation; and the trust wording in the limitations and conclusion is aligned with Section 7's distinction. The workflow statement now requires explicit human responsibility; the evidence requirement is scoped to the categories that apply; a venue-compatible alternative to the byline mark is authorized; and the exposure claim in Section 8 is

stated precisely. The service count is unchanged at five; ChatGPT's tags add d, recording that the conformance section's first draft arose from its review exchange.

Version 1.7 (deposited 29 July 2026) — changes from version 1.6.3. One correction, one specification change, and a review pass. Correction: a production defect in version 1.6.3 caused one sentence to appear five times inside Section 10 (Limitations), interrupting two sentences; the intended text is restored. The defect arose when the document was built, not from an edit to the argument, and was found in independent adversarial review — recorded in the AI-use disclosure below, which also revises one service's function tags from (s) to (c/s), a revision that changes no count. The disclosed service count does rise, from four to five, because a fifth service was used as an adversarial reviewer before this deposit; the byline reads LI-AIast5 accordingly, and the higher number records additional use, not additional merit. Specification: the canonical code table in Section 3.1 is headed **Service** and lists only services, applying the rule already stated in that section that the service is the unit the author used; open-weight models with no first-party service are disclosed by the serving platform, and any service not listed is written out in full. Wording throughout is aligned to the same terminology. Review pass: successive independent reviews before deposit produced these further corrections — Section 6's worked examples are labeled A–D and the challenged-disclosure example now matches the tags it refers to; the claim that every abuse is detectable is corrected, since omission cannot be revealed by produced evidence; "sole author of record" is reworded to remove a collision with multi-author works; the ICMJE recommendations are cited to their current January 2026 edition with the cover-letter requirement stated; the STM classification summary enumerates all nine activity categories; one preprint is cited to its named author and to the version whose title it carries; one reference moved from advance online publication to its assigned volume and page range; the canonical code list in the Glossary is aligned to the Section 3.1 table and the notation is written SERVICE(fn); each worked example now carries the workflow and evidence statement the standard requires; the claim that a structured mark satisfies a journal mandate is bounded by each journal's placement and cover-letter rules; the claim that misuse of the LI prefix costs a reader nothing is narrowed; structured disclosure is described as supporting comparison rather than ranking, and the evidence a claim rests on as retained evidence rather than transcripts alone; spelling is standardized to US forms; and several sentence-level errors are fixed. No part of the argument or the function tags has changed.

References

Allen, L., Scott, J., Brand, A., Hlava, M., & Altman, M. (2014). Publishing: Credit where credit is due. *Nature, 508,* 312–313. https://doi.org/10.1038/508312a.

Brand, A., Allen, L., Altman, M., Hlava, M., & Scott, J. (2015). Beyond authorship: attribution, contribution, collaboration, and credit. *Learned Publishing, 28*(2), 151–155. https://doi.org/10.1087/20150211.

Cleland, J., Driessen, E., Masters, K., Lingard, L., & Maggio, L. A. (2026). When and how to disclose AI use in academic publishing: AMEE Guide No. 192. *Medical Teacher,* 48(4), 542–553. https://doi.org/10.1080/0142159X.2025.2607513.

Glynn, A. (2025). *Academ-AI: documenting the undisclosed use of generative artificial intelligence in academic publishing* (Version 2, 15 November 2025). arXiv:2411.15218.

International Committee of Medical Journal Editors (ICMJE). (2026). *Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals* (updated January 2026; Section V, Use of Artificial Intelligence in Publishing — AI-assisted technology guidance introduced May 2023). https://www.icmje.org/icmje-recommendations.pdf.

International Science Council, COPE, STM, and the Global Young Academy. (2026). *Global Reporting Standard for AI Disclosure in Research* — public consultation (with the World Conferences on Research Integrity Foundation supporting the Vancouver focus track). https://council.science/our-work/ai-disclosure-in-research/

National Information Standards Organization (NISO). (2022). *ANSI/NISO Z39.104-2022, CRediT (Contributor Roles Taxonomy).* https://doi.org/10.3789/ansi.niso.z39.104-2022.

Nature Portfolio. (2024). *Editorial policies: Artificial Intelligence (AI).* https://www.nature.com/nature-portfolio/editorial-policies/ai (continuously updated; accessed July 26, 2026).

Resnik, D. B., & Hosseini, M. (2026). Disclosing artificial intelligence use in scientific research and publication: When should disclosure be mandatory, optional, or unnecessary? *Accountability in Research*, 33(2), Article 2481949. https://doi.org/10.1080/08989621.2025.2481949.

Schilke, O., & Reimann, M. (2025). The transparency dilemma: How AI disclosure erodes trust. *Organizational Behavior and Human Decision Processes*, 188, 104405. https://doi.org/10.1016/j.obhdp.2025.104405.

STM (International Association of Scientific, Technical and Medical Publishers). (2025). *Recommendations for a Classification of AI Use in Academic Manuscript Preparation.* 19 September 2025. https://stm-assoc.org/document/recommendations-for-a-classification-of-ai-use-in-academic-manuscript-preparation/.

Suchikova, Y., Tsybuliak, N., Teixeira da Silva, J. A., & Nazarovets, S. (2026). GAIDeT (Generative AI Delegation Taxonomy): A taxonomy for humans to delegate tasks to generative artificial intelligence in scientific research and publishing. *Accountability in Research*, 33(3), Article 2544331. https://doi.org/10.1080/08989621.2025.2544331.

Weaver, K. D. (2024). The Artificial Intelligence Disclosure (AID) Framework: An introduction. *College & Research Libraries News, 85*(10), 407–411. https://doi.org/10.5860/crln.85.10.407.

Weaver, K. D. (2026). Artificial intelligence and transparency: Toward a framework for disclosure of AI use in learning, research, and publication. *Information Services and Use*, 46(1–2), 5–11. https://doi.org/10.1177/18758789261435716.

Young, J. (2026). AIR: AI in Research — a framework for transparent and responsible AI use mapped to the research process. *Figshare* (online resource). https://doi.org/10.6084/m9.figshare.31268020.