

Labelling the Entanglement: The AI Usage Facts Tool as Reflective Practice, Self-Disclosure, and Epistemic Infrastructure

Faraz Forghan Parast

Electronic Textual Cultures Lab, University of Victoria

Abstract

Generative AI increasingly enters scholarly writing as a cognitive and linguistic prosthetic: a tool that extends thought through language while hiding its mediation in fluency. This article introduces the AI Usage Facts label (<https://ailabel.netlify.app>), a customizable disclosure tool modelled on the nutrition facts label, and argues that disclosure can become a practice of self-knowledge rather than a ritual of compliance. The label asks authors to reconstruct where AI entered the work, how heavily each stage relied on it, and what forms of verification followed. In doing so, it creates deliberate friction: a structured pause in which the writer's entanglement with the model becomes perceptible again. The article shows how the label's design choices, including staged workflow reconstruction, Human/Cyborg/AI classifications, a rough percentage index, quality attestations, and live preview, translate a phenomenological account of prosthetic authorship into an interface. It also argues that granular disclosure alters the affective economy of AI use. A detailed account of assistance, reliance, and verification offers readers something more useful than the bare confession "I used AI": a basis for calibrated judgment. The article closes by addressing the limits of self-report and the possibility of gaming, arguing that the label cannot guarantee honesty but can cultivate a habit of attention. What begins as a disclosure form may become a small piece of epistemic infrastructure for scholarly writing in the age of cognitive prosthetics.

This article is the second panel in a three-part sequence on generative AI, prosthetic authorship, and scholarly literacy. The first article develops the phenomenological account of generative AI as a cognitive and linguistic prosthetic, focusing on fluency, friction, and premature transparency. The present article turns from diagnosis to disclosure infrastructure through the AI Usage Facts label, a tool designed to make AI mediation perceptible within scholarly workflows. A third article develops a pedagogical framework for prosthetic literacy. Each article is designed to stand alone, but together they trace a movement from phenomenological diagnosis to practical disclosure to educational rehabilitation.

Introduction: From Prosthesis to Label

The first article in this sequence argued that generative AI functions as a cognitive and linguistic prosthetic: a mediating apparatus that extends, redistributes, and partially reorganizes the labour of thought and expression, a continuing question posed by work on extended and distributed cognition (Clark & Chalmers, 1998; Hutchins, 1995; Clark, 2003). It developed the concept of premature transparency to name a condition in which a technology becomes functionally incorporated before it becomes reflectively recognized as mediation. Large language models enter through the most intimate medium scholars possess (language itself), and their outputs can dissolve into the ordinary flow of composition before the user has learned to perceive them as anything other than their own developing thought. The phenomenological argument concluded that responsible AI literacy requires practices of re-sensitization: deliberate interruptions through which the user pauses to reconstruct where cognitive labour has been routed, transformed, or delegated.

The present article takes up that problem at the level of design. The AI Usage Facts label is the material continuation of the phenomenological argument: the point at which an account of prosthetic authorship meets the practical question of what to build, and how the thing built can itself become a practice of self-knowledge.

The AI Usage Facts label maker, available at <http://ailabel.netlify.app>, is a web-based disclosure tool that invites authors to reconstruct the cognitive trajectory of AI involvement across the stages of scholarly research and writing. It generates a fact sheet modelled on the nutrition facts label available on food packages, providing a structured framework for articulating which tasks involved AI assistance, to what degree, under what quality controls, and with what tools. In practice, the interface works as a reflective scaffold. Each stage of the workflow becomes an occasion for the author to pause and interrogate a process that fluent AI mediation tends to render invisible. The live preview updates as decisions are made, creating a continuous visual representation of one's own distributed cognitive labour. The tool also accommodates disciplinary plurality through customizable workflow stages, so that a historian's reflective categories need not be forced into the vocabulary of a molecular biologist.

The argument that follows moves through each of these functions in turn to show how the design's individual choices encode commitments about what disclosure should do to the person performing it.

The Nutrition-Label Metaphor: Informing Without Moralizing

The choice of metaphor is deliberate. Nutrition labels emerged in the mid-twentieth century in response to a very specific epistemic issue: consumers could not determine what was in their food by looking at it. Processed food was often beneficial, serving legitimate purposes of preservation, convenience, and nutrition. But the gap between appearance and composition made informed choice impossible. The label standardized disclosure in a recognizable, clear format without banning processed food or moralizing about its consumption. It quantified components that had been vague, itemized ingredients, warned of potential allergens, and returned to the consumer the information necessary for evaluation. The label's cultural message was as important as its informational content: processing is widespread, often beneficial, and deserves informed evaluation rather than moral panic.

The nutrition label's history is interesting in ways that extend beyond this metaphorical aspect. The label as it now exists in North America was standardized under the Nutrition Labeling and Education Act of 1990 and implemented in 1994, in response to a problem of information architecture: manufacturers were presenting nutritional data in inconsistent formats, with wildly varying serving sizes and selectively highlighted figures, making comparison across products practically impossible. Consumers could not evaluate what they were eating not because the information was absent but because it was presented in a register designed to serve marketing rather than comprehension. The regulatory response was not to ban processed food or require confession. It was to create a standardized vocabulary for composition, rendered in a deliberately neutral visual register — black and white, fixed typography, no brand colours — that signalled its separation from the promotional space surrounding it. The front of a package is marketing. The nutrition label belongs to a different epistemic register entirely, one that announced: this information is not trying to sell you anything.

The percentage daily value column was an innovation of that moment. Before standardization, a consumer who wanted to evaluate sodium content had no ready framework for calibrating whether 480mg was high or low. The daily value percentage created a comparative infrastructure: a shared denominator that allowed evaluation across products, serving sizes, and dietary contexts. The label did not tell consumers what to eat. It created the conditions under which informed choice became thinkable.

The 2016 FDA revision (U.S. Food and Drug Administration, 2016) is equally informative, because it reveals that the label was never a finished instrument. Updated serving sizes were recalibrated to reflect what people actually eat rather than idealized portions. Added sugars received their own line, separated from naturally occurring sugars, because the scientific consensus on free sugars had developed in the intervening decades and the original label had obscured a distinction that now mattered. The calorie count was given larger type. These revisions responded to evidence about how the label was being read, misread, and circumvented. The label is an evolving framework, not a static disclosure, and that evolutionary responsiveness is itself part of what makes it infrastructure rather than packaging.

The AI Usage Facts label follows this developmental and aesthetic logic. It responds to a gap between the polished surface of a scholarly manuscript and the conditions of its production — a gap that, like the gap between a food product's appearance and its composition, cannot be closed by looking at the surface alone. It adopts a neutral, structured format deliberately separated from the manuscript's rhetorical register. It creates a shared vocabulary — Human, Cyborg, AI; stage-level classifications; quality attestations — within which comparison and calibration become possible across authors, disciplines, and institutions. And it is designed, from the outset, to be revisable: the customizable stage model means that as disciplinary communities develop more nuanced understandings of where cognitive labour matters most, the framework can evolve without abandoning the underlying structure. The nutrition label took decades to revise. The AI label is designed to be revised more quickly, because the technology it discloses is moving faster than food processing ever did.

The AI Usage Facts label applies this logic to scholarly texts. AI-assisted writing, like processed food, is often careful, proportionate, and well-governed. But the gap between the polished surface of a finished manuscript and the conditions under which it was produced has widened beyond what existing scholarly conventions can bridge. Citations, method sections, acknowledgements, and ethics statements (the traditional apparatus of academic provenance) were designed for a world in which coherent prose reliably indicated human cognitive labour. That reliability has been disrupted. The label responds by adding an explicit layer of process metadata where implicit signals of provenance have become insufficient.

The metaphor also carries a particular affective register. A nutrition label describes composition in neutral, quantitative terms and trusts the consumer to form their own judgement, without asking for confession or dividing food into pure and contaminated categories. This affective neutrality is deliberate in the AI context. Current discourse around AI-assisted scholarship is saturated with moral binaries: human versus machine, authentic versus fake, earned versus unearned, scholarship versus "slop." These binaries flatten the gradient of involvement that characterizes actual practice and produce a climate in which any AI involvement risks being read as total contamination. The nutrition-label format intervenes in this economy by normalizing description over confession. It replaces "I used AI" (a statement that sticks, in Ahmed's (2004) sense, to shame and suspicion) with a situated account of where AI entered, how heavily the workflow leaned on it, and what verification followed. The format invites interpretation rather than verdict.

Nutrition labels are also infrastructural. They created a shared vocabulary for discussing food composition, enabled comparative evaluation across products, supported regulatory oversight, and generated data that researchers could analyse at scale (U.S. Food and Drug Administration, 2020). The AI Usage Facts label aspires to a parallel transformation in scholarly culture: from vague binary judgements about AI involvement toward a more textured, evidence-based understanding of how human and machine labour are distributed in the production of knowledge.

Design as Philosophy: The Architecture of Reflective Encounter

Design, here, becomes argument in interactive form. In this sense, it belongs to a postphenomenological understanding of technologies as mediators of perception, action, and self-relation rather than neutral instruments (Ihde, 1990; Verbeek, 2005). Each element, including workflow reconstruction, staged classification, the quantitative index, quality attestations, and the live mirror of one's emerging self-description, embodies a philosophical commitment about what disclosure should accomplish and what kind of cognitive experience it should occasion. The design philosophy draws from the phenomenological account developed in Part 1: if AI mediation becomes transparent too quickly, sedimenting into the writer's cognitive routine before it has been reflectively recognized, then the instrument of disclosure must create conditions under which that mediation becomes temporarily opaque again and available for inspection.

Reconstructing the Stages of Cognitive Labour

The first reflective encounter the label occasions is the act of decomposing one's work into stages, recognizing that scholarly production is a sequence of distinguishable moments, each with its own epistemic character and its own vulnerability to prosthetic mediation. Default stage models based on established taxonomies (notably CRediT, the Contributor Roles Taxonomy) offer a vocabulary for self-observation, a set of named moments against which the author can begin to map the actual trajectory of their cognitive labour.

The customizability of these stages serves two distinguishable cognitive purposes. The first is representational adequacy: different forms of scholarship have different structures, and a reflective instrument that forces all disciplines into a single template would be too generic to occasion meaningful self-recognition. A creative writer might define stages as ideation, drafting, revision, and line editing. An empirical scientist might distinguish hypothesis formation, experimental design, data collection, and manuscript preparation. A historian might include archival research, source analysis, and narrative construction. Each taxonomy represents a disciplinary community's understanding of where epistemic weight resides. Negotiating that taxonomy is itself a form of collective self-knowledge.

The second purpose is pedagogical, operating through the phenomenology of decomposition itself. When assigned as a component of research training, naming stages becomes a teaching instrument. Students learn to see their writing process as articulated rather than monolithic. They recognize that "writing a paper" dissolves, under reflective attention, into distinguishable cognitive acts, each of which may have been differently mediated. The experience of decomposing one's workflow is itself a practice of prosthetic literacy, because it requires inhabiting the question: where did the thinking happen?

The Cognitive Work of Classification: Human, Cyborg, AI

Within each stage, individual tasks are classified along a scale rendered through three designations: Human, Cyborg, and AI. The terminology is significant: "Cyborg" (drawn from Haraway's (1991) theorization of hybrid human-machine existence) describes the middle ground

that binary discourse about AI authorship systematically erases. Most AI-assisted scholarship is cyborg work; that is, thought that has been scaffolded, extended, or partially reorganized through interaction with a language model. The three-category classification makes this middle ground speakable and cognitively prominent, refusing the collapsed dichotomy that forces every moment of mediated work into either "mine" or "the machine's."

Finer-grained scales (a five-point option introducing quarter and three-quarter increments) are available for those who wish to capture more nuance. But the default three-point scale carries deliberate cognitive weight. It asks the author to make a primary determination: was this done by the human, done through collaboration, or done by the AI? That determination, however rough, is itself a moment of re-sensitization. It interrupts the seamless retrospective sense that "I wrote this" and requires the author to consider whether, more accurately, "we wrote this" or "it drafted this and the revision was mine."

What matters phenomenologically is the cognitive work the classification demands. The moment of hesitation (did the model's suggested phrasing merely polish what was already there, or did it redirect the argument's trajectory?) is the moment in which premature transparency is temporarily reversed. The author who pauses over this distinction is perceiving the prosthetic relation rather than simply inhabiting it. Each act of classification is the trace of a reflective act in which the author confronts the distributed character of their own cognitive process.

Quantification as Cognitive Mirror: The Percentage Index

The percentage index presents a single figure ("Total AI Contribution: 23%"), derived from task-level classifications, weighted by the epistemic significance of each stage. Critics might object that authorship cannot be quantified, or that any such metric will be gamed or misunderstood. These objections are important, but they misidentify the function of the index. The percentage does not claim to measure authorship in an ontological sense. Such a claim would contradict the prosthetic framework, where the very notion of cleanly separable contributions dissolves under phenomenological scrutiny. The index estimates delegated cognitive load: the degree to which tasks were offloaded to, scaffolded by, or materially shaped through AI systems. It is a measure of reliance.

The index's most important function may be experiential rather than communicative. The number operates as a cognitive mirror: a condensed representation of one's distributed process that often

surprises. The author who assumed their AI use was minimal may discover, through accumulated classifications, that the aggregate tells a different story. Conversely, the scholar who harboured anxiety about over-reliance may find reassurance in seeing that the core intellectual labour remained decisively human. In both cases, the number performs a calibrating function, intervening between the author's vague self-impression and a more articulate account. The index is an invitation to closer inspection and conversation.

Verification as Epistemic Commitment: Quality Attestations

The label includes a quality attestation dimension prompting the author to articulate which verification practices accompanied their use of AI: human verification of claims, plagiarism scanning, citation cross-checking, assessment of data leakage risk, and ethics review where applicable. The cognitive significance lies in the self-interrogation these attestations occasion. To attest that citations were cross-checked is to have asked oneself whether they were. To leave an attestation unmarked is to confront a gap in one's epistemic care, a moment of uncomfortable self-knowledge that may prompt future diligence.

The quality dimension reframes evaluation from “how much AI was used?” to “how was AI use governed?”, a shift especially important given well-documented risks of automation bias, over-reliance, and miscalibrated trust in automated systems (Parasuraman & Manzey, 2010; Lucas et al., 2024). Two manuscripts may both reflect substantial AI involvement, yet they are epistemically different objects if one has undergone rigorous verification while the other has not. The reframing honours the cognitive labour of verification rather than treating all AI involvement as epistemically equivalent. Standardized warnings about residual risks (hallucinated claims, training-cutoff gaps, stylistic biases) serve a complementary function, embedding epistemic humility into the disclosure practice and acknowledging what was described in part 1 as the irreducible opacity of generative systems.

Deliberate Friction: Disclosure as Re-Sensitization

At first glance, the label's structure (staged categories, classification scales, percentage summaries, attestation prompts) might seem closer to institutional audit than to phenomenological inquiry. But the specificity of the reflective structure is what transforms vague self-impression into accountable self-knowledge, using friction not as bureaucratic drag but as a

metacognitive cue that interrupts the false ease of fluent processing (Alter & Oppenheimer, 2009). To engage with the label, the author must pause over each stage and ask: did AI enter here? What did it change? How was its output evaluated? The pause is the label's most important feature, easy to miss because it is enacted through the seemingly mundane practice of classifying recalled experience.

Part 1 developed the concept of premature transparency to describe a specific problem: LLMs can become incorporated into the writer's cognitive apparatus before the user has learned to perceive them as mediation. Where visible prosthetics (carbon-fiber blades, cochlear implants) make mediation perceptible through material difference, language models cloak their mediation in fluency. The label's deliberate friction interrupts that fluency, creating through structured reflective practice a moment of perceptibility for a prosthesis whose mediation is otherwise hidden in linguistic ease. Each act of classification is an exercise in what part 1 called the reversal of sedimented coupling: a moment in which the tool that had become part of the cognitive background is brought into the foreground of attention.

Consider a historian completing the label after drafting an article with intermittent AI assistance. When the reconstruction reaches the stage of source analysis, a hesitation arises. The model did not invent the archival evidence, but it did suggest a relation between two documents that later shaped the argument's trajectory. Was that contribution ideational, synthetic, or interpretive? Should the task be classified as Human, Cyborg, or AI? The label cannot answer these questions, yet it makes them unavoidable.

Or consider a graduate student who prompted a language model for help structuring an argument, then revised extensively, introducing new evidence and reordering claims. When the label asks for a classification of the argumentation stage, the student faces a difficult question: was this scaffolded thinking or generated thinking?

The live mirror of the emerging label reinforces this dynamic. As reflective decisions accumulate, the visual representation updates in real time. Stage summaries shift; the total index recalculates. The immediacy gives the author a continuous external representation of their own workflow. The preview functions as a mirror of the process through which the text was made, rather than of the finished text itself. For many authors, watching this take shape (seeing that AI contributed more to the literature review than casual recollection suggested, or less to the

argumentative structure than anxiety had feared) is itself a moment of cognitive calibration. The label teaches the author about their own practice, performing the reversal of premature transparency.

The structured classification functions like a rehabilitation exercise: a protocol through which authors learn to notice, name, and evaluate the supports through which their work has been produced. A person learning to move with a prosthetic limb does not simply identify where the body ends and the device begins. They learn the altered feel of action itself. The author learning to write with a cognitive prosthetic faces a parallel task: not to recover a pure boundary between self and tool, but to become more articulate about the relation through which the work took shape.

Self-Disclosure and the Affective Economy of AI Use

Disclosure of AI involvement in scholarly work is never neutral. It takes place within an affective economy, a field of circulating intensities in which certain practices have become sticky with suspicion, shame, and moral judgement. Ahmed's (2004) concept of affective economies illuminates the mechanism: affects circulate through social fields, adhering to signs and practices in ways that shape what feels possible or dangerous. In contemporary academic culture, "AI-assisted" has become sticky with fraud, laziness, and cognitive cheapness (Chan, 2024; Lund et al., 2025), a suspicion intensified by emerging debates over AI-assisted writing and academic misconduct, while "human-authored" retains its association with authenticity and intellectual merit.

An affective landscape of this kind creates a disclosure dilemma. Goffman's (1963) work on stigma management illuminates the structure: AI use increasingly occupies the position of a concealable, potentially discreditable attribute. A scholar who has used AI carefully may still anticipate that disclosure will trigger judgement disproportionate to the actual practice, that "some AI involvement" will be read as "total AI generation" in the compressed moral categories of the current moment. Under these conditions, many scholars choose concealment, even when their use was careful, partial, and well-governed.

Scholarly practices are already read through such affective filters. A citation may be received as intellectual honesty, disciplinary positioning, or self-importance. An acknowledgement may appear generous or strategic. A disclosure of AI assistance enters this same interpretive field,

where small paratextual signals are quickly moralized. The problem is therefore not simply how to transmit information about AI use, but how to shape the conditions under which that information will be read. A circular economy of suspicion takes hold: stigma encourages non-disclosure, non-disclosure intensifies suspicion, and suspicion strengthens stigma (Link & Phelan, 2001; Chaudoir & Fisher, 2010).

The AI Usage Facts label intervenes through granularity. By replacing the binary "I used AI" with a structured account of where AI entered, how much reliance it carried, and what verification followed, the label changes the object of judgement. A disclosure indicating 12% AI contribution, primarily in literature search and prose polishing, with outputs verified and citations cross-checked, creates a different affective object than the unstructured confession "I used ChatGPT." The structured account invites proportional evaluation; the bare confession invites binary judgement.

Granularity also serves the author's self-relation. Many scholars who use AI experience anxiety about whether their work is "really theirs." The label provides reassurance through articulation, making visible the cognitive labour that remained human. By reconstructing where AI entered, the author can perceive that the core intellectual work (the research question, the interpretive framework, the argumentative structure) was their own. The label restores proportion and agency to a practice that unstructured anxiety has distorted, functioning as a counter-practice to the affective opacity surrounding prosthetic authorship.

The affective design of the reflective experience supports this reorientation. A disclosure practice that feels punitive will intensify shame and defensiveness. The label pairs deliberate reflection with a responsive structure: stages can be revisited, classifications revised as self-understanding sharpens, and the workflow framework adapted rather than imposed. The goal is to make disclosure's friction feel purposeful, to create conditions under which slowing down feels like self-knowledge.

Pedagogy: Teaching Prosthetic Literacy Through Structured Self-Observation

The label's pedagogical function may be its most consequential. Existing approaches to teaching AI literacy tend toward either policy-driven compliance or skill-driven competence, reflecting

broader efforts to define AI literacy as a set of conceptual, ethical, and practical competencies (Long & Magerko, 2020; Ng et al., 2021; Kasneci et al., 2023). Both have value, but neither fully cultivates prosthetic literacy: the capacity for self-awareness about where cognitive labour has been distributed and how the prosthetic relation has shaped one's thinking. The third article in this sequence develops that pedagogical problem more fully. The present concern is narrower: how a disclosure tool can teach prosthetic literacy by making self-observation repeatable.

The AI Usage Facts label addresses this gap by making self-awareness a structured, repeatable practice rather than an aspirational virtue. When a student completes the label after drafting a research paper, the exercise does more than document prior conduct. The student reconstructs their workflow stage by stage, confronting at each point the question of how their thinking was scaffolded, extended, or partially produced. The reconstruction develops the capacity to perceive one's own mediated process, which is the very capacity that premature transparency degrades. The student who has paused to classify each stage has practiced, in a structured way, the reflective attention that responsible prosthetic authorship demands.

Faculty can design stage models tailored to their assignments, foregrounding the cognitive activities they wish students to attend to. An instructor teaching historiography might create stages emphasizing source evaluation and interpretive reasoning, directing reflective attention to where historical thinking is most distinctive and most vulnerable to prosthetic substitution. A writing instructor might distinguish ideation from drafting from revision, helping students recognize that these activities involve different cognitive demands and different risks of what was referred to in part 1 as generative opacity (the condition in which AI's contribution is difficult to disentangle because it entered at a generative rather than editorial level). Mollick and Mollick's (2023) taxonomy of AI integration (AI as tutor, coach, collaborator, mentor, teammate) provides a complementary frame: where their taxonomy describes roles AI can play, the label provides a reflective instrument for reconstructing which roles AI actually played.

The temporal dimension is critical. Disclosure anchored entirely in retrospection is structurally fragile because it asks memory to perform source reconstruction that fluent AI assistance degrades. The very condition of premature transparency means cognitive traces of mediation fade faster than for conspicuous tools. The label is most effective when engaged close to composition, as concurrent self-observation. Faculty can invite students to update their

reconstruction at the end of each drafting session, preserving the phenomenological freshness. The author is still close enough to recall where the model's contribution began and where their own thinking reasserted itself.

The approach transforms disclosure from compliance into learning: "What did you notice about your own process?" rather than "Did you follow the rules?" Over a semester, students who engage with the label for multiple assignments can develop a longitudinal record of their evolving relationship with AI, where they relied on it more or less, where verification improved, where they learned to do without it. The record is prosthetic literacy in practice: cultivated self-awareness about one's own prosthetic condition, developed through the repeated practice of pausing to articulate how one's work was made.

Epistemic Infrastructure: From Individual Reflection to Collective Self-Knowledge

Individual reflection, however valuable, remains private and ephemeral unless captured in a form that can travel across documents, institutions, disciplines, and time. The AI Usage Facts label produces an artifact of self-observation: a standardized account that can be appended to manuscripts, deposited in repositories, indexed by search engines, and analysed at scale.

The label can be rendered as a human-readable visual for supplementary materials and as structured machine-readable metadata for repositories, indexing services, or future verification tools. The machine-readable format opens possibilities extending beyond individual disclosure: large-scale analysis of AI use patterns, automated assessment of disclosure completeness, integration with ORCID, DOI systems, or institutional repositories. What begins as a private act of reflective reconstruction can thus accumulate into a collective record of how cognitive prosthetics are reshaping scholarly practice.

Even if every scholar disclosed honestly, without a shared framework those disclosures would remain incommensurable. One author might write "AI was used for light editing"; another "ChatGPT assisted with literature review and drafting." The label standardizes the vocabulary and structure of disclosure without standardizing its content. Authors retain full control over what they disclose; the framework provides the grammar within which disclosure becomes communicable, comparable, and cumulative.

Looking forward, journals could incorporate such disclosures in submission processes. Institutions could use aggregated data to develop evidence-based policies. Disciplines could develop their own stage models and epistemic weights, a process of norm-negotiation that is itself epistemically productive. The label adds another layer to academic provenance, alongside citations, DOIs, CRediT roles, ORCID identifiers, and ethics statements, one addressing the specific gap created by cognitive prosthetics whose mediation is invisible in the finished text (Brand et al., 2015; ORCID, 2022), extending scholarly provenance into the cognitive process by which human and machine labour were entangled in producing the text itself.

These possibilities describe a developmental arc that voluntary scholarly disclosure instruments have followed before, and that the AI Usage Facts label is designed, from the level of its architecture, to be capable of traveling.

The Contributor Roles Taxonomy (CRediT) is the most instructive example. Developed collaboratively between 2012 and 2014 by a small group of journals, publishers, and scholarly communication researchers, CRediT standardized the vocabulary for describing what contributors to a scholarly work actually did: conceptualization, methodology, software development, formal analysis, writing (original draft), writing (review and editing), supervision, funding acquisition, and so on. Its purpose was to address a recognized inadequacy in author order as a proxy for contribution — an inadequacy that had become more acute as research teams grew larger and as the diversity of scholarly labour made single-author conventions meaningless. CRediT offered no enforcement mechanism. Its value was entirely dependent on voluntary adoption by authors and journals. Within a decade, it had been incorporated into the submission systems of major publishers including Elsevier, Springer Nature, and Wiley. It is now approaching de facto standard status in biomedical and many social science fields, with some journals moving from voluntary to required disclosure. The arc from speculative vocabulary to embedded infrastructure took approximately ten years.

ORCID (Open Researcher and Contributor ID) followed a similar trajectory. Launched in 2012 as a voluntary persistent identifier system for researchers, it had no intrinsic authority. Researchers could register or not. Its utility was entirely dependent on network effects: the identifier was only valuable if enough systems recognized it. Funders changed the dynamic. The National Institutes of Health, the Wellcome Trust, and UK Research and Innovation began

requiring ORCID iDs in grant applications, effectively mandating participation for grant-seeking researchers. Publishers followed. ORCID is now structurally embedded in scholarly communication infrastructure in ways that its founders would have been cautious to predict in 2012, precisely because such predictions would have sounded like overclaiming about an instrument that was, at the time, a promising voluntary system.

Ethics statements in human-subjects research offer a third case. For decades, disclosure of ethical review and informed consent processes was field-specific, voluntary in many contexts, and inconsistently applied. The journey toward standardization was driven by the accumulation of journal policies, funder requirements, and disciplinary norms, each reinforcing the others until the statement became expected infrastructure. If an author were to submit to a biomedical journal without an ethics statement in the present day, they would not simply fail a compliance check. They would be signalling unfamiliarity with the basic conventions of the field.

What these precedents share is a structural sequence: a recognized gap in scholarly communication provenance, a standardized voluntary framework offering a shared vocabulary and format, early adoption by high-profile journals or funders, network effects that make adoption increasingly the path of least resistance, and eventual incorporation into submission infrastructure.

The claim here is that the label is designed to be capable of following this arc, one that has been traveled before by instruments that began as voluntary reflective practices and became embedded scholarly infrastructure. The distance between the nutrition label as metaphor and the AI Usage Facts label as institutional precedent in its own right is shorter than it first appears. Both are disclosure instruments designed to address a gap between surface appearance and underlying composition. Both standardize vocabulary in a neutral register separated from promotional framing, and tell consumers or scholars to create the conditions under which informed judgement becomes possible. The nutrition label took three decades to move from first implementation to its 2016 revision, during which time gaming strategies were identified, loopholes were closed, and serving size definitions were updated to reflect actual consumption. The AI label will face its own equivalent pressures. The appropriate response, rather than treating those pressures as objections to the instrument's legitimacy, is to design the instrument for revisability. In the form of its customizable architecture and open metadata format, it already is.

The Self-Report Problem and Its Productive Limitations

An honest account must address the label's most obvious limitation: dependence on self-report. The label asks authors to reconstruct their workflow and classify their own AI use. The reconstruction is imperfect, especially when use has been iterative, habitual, or partially remembered. Source-monitoring research (Johnson et al., 1993) confirms that humans are poor at tracking origins of information arriving in fluent, contextually appropriate form (the very form in which LLM outputs appear). The condition making the label necessary (the tendency of fluent mediation to become cognitively transparent) also makes perfect retrospective reconstruction impossible.

The limitation does not, however, nullify the reflective practice. The strongest use of the label is concurrent or near-concurrent self-observation. Even retrospectively, the structured framework converts vague impressions into specific classifications, producing a more articulate record than would otherwise exist. What the label yields is estimated cognitive distribution: a disciplined attempt to describe how human and AI labour were entangled in making a text.

Imperfect self-report has a productive dimension connecting directly to the phenomenological account. The moments where the author hesitates, unable to determine whether a contribution was "theirs" or "the model's", are disclosures in their own right. The hesitation is a phenomenological datum: evidence that prosthetic coupling has reached a depth where clean separation is no longer possible. An author who classifies a task as Cyborg because contributions cannot be untangled provides more epistemically valuable information than one who confidently assigns clean categories. The Cyborg classification exists to honour this productive uncertainty, to name the entanglement at the heart of prosthetic authorship.

Gaming, the pause, and reader relief

A further risk is that disclosure becomes a hollow compliance ritual. This risk is shared with many scholarly conventions: citations, methods sections, and ethics statements can all become formulaic. Scholarship retains them because even imperfect conventions still create friction, accountability, and shared expectations. The answer lies in designing disclosure practices that remain connected to pedagogy and interpretive engagement, practices resisting routinization by continuing to motivate true cognitive work.

The nutrition-label analogy, however, carries a complication the paper must address directly. Nutrition labels have been gamed methodically, and the strategies used illuminate both the risks facing the AI disclosure instrument and the structural features that partially protect against them. One common manipulation works by shrinking the unit of disclosure below the unit of actual consumption. A premium ice cream manufacturer may define a serving as 30 grams; that is a quantity no person consuming ice cream from a tub will ever measure out or observe. The calorie, fat, and sugar figures for that serving are accurate. The label is not false. But the serving size is designed so that the figures feel innocuous, while the calories consumed in a normal bowl are two or three times what the label prominently displays. A related strategy exploits regulatory thresholds: in the United States, a product may claim zero grams of trans fat per serving if the actual content falls below 0.5 grams. A coffee creamer manufacturer who sets the serving size at one teaspoon or one gram can therefore advertise "0g trans fat" while delivering approximately 0.75 grams per serving — a figure that, consumed across three or four additions to a single morning's coffee, accumulates to a medically significant dose. The label is legally compliant and practically misleading. The loophole was found, engineered around, and put to work.

A third strategy operates through fragmentation. Ingredient lists must be ordered by weight, with the heaviest ingredient first. A granola bar manufacturer who "sugar-splits" lists cane sugar, brown sugar, molasses, honey, maltodextrin, and glucose syrup as separate ingredients, thus ensuring that none of them appears near the top of the list individually, even if their combined weight would place them first. "Whole rolled oats" leads. The consumer reads a natural, wholesome product. The fragmentation is designed to exploit the ordering rule by disaggregating what is, nutritionally, a single category.

These strategies share a common architecture: they all work by manipulating the granularity of disclosure so that the unit reported is smaller than the unit that matters. The serving size is smaller than what you eat. The per-serving value is smaller than what you ingest. The individual sugar is smaller than the total sugar load. The label is technically complete while being practically deceptive.

The question for any AI disclosure instrument is whether analogous strategies are available and likely. They are. An author could classify AI-generated structural scaffolding as Human because they typed the final sentences themselves. They could use the Cyborg category as a catch-all that

obscures wholesale generation at key stages. They could describe extensive AI involvement in argumentation as "light editing assistance" by selecting whichever stage label fits the most defensible reading of what happened. These moves are available, and a sufficiently motivated actor will use them.

What the AI Usage Facts label's design offers, however, is a structural feature that nutrition label gaming did not have to contend with: the primary subject of the disclosure is also its author. Nutrition label manipulation happens upstream, in the design of the label before the consumer encounters it. The consumer who reads the label in good faith is still deceived, because the architecture was already rigged. The AI label works differently. The author completing the label is not reading someone else's claims about a product they are selling. They are constructing a self-account, stage by stage, about cognitive labour they performed and can remember. The manipulation, if it occurs, must happen in real time, against the non-trivial resistance of actual memory.

To systematically misclassify a workflow, an author must recall what actually happened, decide to misrepresent it, construct a plausible alternative narrative, and sustain that narrative through a sequence of specific, interrelated classification decisions. The label's deliberate friction — the same friction that makes honest completion an act of re-sensitization — also makes dishonest completion cognitively costly. The granularity that asks the honest author to pause and reflect asks the dishonest author to actively fabricate across multiple stages. This is not a guarantee against gaming. It is a structural asymmetry that makes casual dishonesty more effortful than honest completion, inverting the usual dynamic in which compliance is the path of least resistance.

The deeper answer to the gaming problem, however, is cultivational rather than regulatory. The label is not designed primarily as a verification instrument. Verification requires an external auditor with access to the original drafts, prompts, and revision history; this is a standard no existing scholarly infrastructure enforces and that would be impractical at scale. The label is designed as a reflective practice, and its value is not primarily located in the accuracy of any single completed instance but in the habit of attention that repeated engagement builds. A scholar who completes the label honestly for a dozen projects over a semester develops a progressively sharper capacity to notice, in real time, where AI is entering their process. The instrument

teaches the sensitivity it records. Gaming the label does not generate that sensitivity. The student who misclassifies every stage may produce a plausible-looking label, but they have also willfully bypassed the cognitive work that makes the label worth completing.

This reframing also clarifies the affective dimension from the reader's side of the exchange. This was addressed primarily from the author's position in the earlier discussion of stigma management. The stigma economy, as analyzed through Ahmed and Goffman, correctly describes the author's vulnerability: disclosure risks triggering disproportionate moral judgement in a cultural climate saturated with binary categories. But the affective experience of reading a detailed, specific AI disclosure is different from the experience of encountering "AI was used in the preparation of this work." The difference is not primarily informational. It is phenomenological.

A reader who encounters a completed AI Usage Facts label — with stage-level classifications, a percentage index, quality attestation checkboxes, and a note that the argumentation stage was classified as Cyborg because the author accepted a structural suggestion from the model but reworked the evidence and conclusion independently — is not encountering a confession. They are encountering evidence of metacognitive competence. The specificity signals: this person was paying attention to their own process. They knew what the model did. They evaluated it. They made choices about what to keep and what to revise. That signal is the opposite of the careless delegation that the stigma economy is actually reacting against.

The reader's likely response is not intensified suspicion but something closer to recalibrated trust and, in many cases, relief. The relief is partly epistemic: a proportionate account of AI involvement gives the reader an actual basis for calibrated evaluation, replacing the binary anxiety of "I do not know how much AI is in this" with a structured account they can assess. But it is also partly social. In an ecology where many scholars use AI without disclosing it, and where that non-disclosure is widely suspected even when unprovable, an author who completes a detailed disclosure is doing something that takes a kind of confidence. The reader who has been quietly wondering about a colleague's prose may find, in a specific and honest disclosure, not confirmation of their worst fears but reassurance that a thoughtful person has already asked the hard questions about this particular piece of work, and found that the answers are proportionate,

defensible, and worth sharing. The granularity itself is reassuring because it demonstrates that there was something to be granular about, and that the author was not afraid to say what it was.

This affective dynamic is, structurally, the inverse of what the stigma model predicts for binary disclosure. "I used AI" is sticky in Ahmed's sense because it offers no texture for calibration and therefore invites binary moral categorization. A detailed label breaks the stickiness by multiplying the objects of judgement until none of them is large enough to carry the full freight of moral anxiety. The reader cannot sustain the "authentic versus fake" binary in the face of a disclosure that reads: "argumentation: 15% AI contribution; literature synthesis: 40% AI contribution; revision: 5% AI contribution; all claims human-verified; citations cross-checked." The binary has been dissolved by specificity before the reader has had a chance to apply it.

Conclusion: Disclosure as Epistemic Virtue

The AI Usage Facts label is a small intervention in a vast and rapidly shifting landscape. It cannot prevent dishonesty, guarantee accuracy, or eliminate the interpretive difficulty of distributed authorship. What it does is create a durable occasion for authorship to become self-descriptive under conditions of technological mediation, a structured practice through which the author's relationship to their cognitive prosthetics is made, however temporarily, an object of reflective attention.

The label's four functions are best understood as one pause extended across scales. For the writer, it interrupts the fluency through which prosthetic mediation disappears into ordinary composition. For the reader, it changes disclosure from a confession into a proportioned account of reliance, verification, and care. For the classroom, it turns that interruption into practice, asking students to learn the altered feel of thinking with a linguistic prosthesis. For scholarly infrastructure, it gives these acts of self-description a portable form, allowing them to travel across documents, repositories, disciplines, and time. What begins as a hesitation before a checkbox may become a grammar for describing mediated authorship.

The design elements carry this philosophy in practical form. The nutrition-label metaphor invites informed evaluation rather than moral panic. Workflow reconstruction acknowledges that different forms of scholarship organize cognitive labour differently, and that naming those differences is already a form of disciplinary self-knowledge. The percentage index offers a rough mirror of reliance without pretending to measure authorial essence. The Human, Cyborg, and AI

classifications make the hybrid middle ground visible and speakable. Quality attestations shift attention from the quantity of AI use to the care with which it was governed. The warnings embed epistemic humility. Structured export allows individual acts of reflection to accumulate into a collective record of how cognitive prosthetics are reshaping scholarly production.

Part 1 argued that generative AI has been deployed as a cognitive prosthetic with essentially no rehabilitation infrastructure. The label is part of that missing infrastructure. It creates a disciplined reflective checkpoint, an external scaffold in Clark's (2003) sense, that can become part of the cognitive routine through which users monitor, distribute, and reappropriate their own thinking. It does so without pretending that distributed cognition can be resolved into a perfectly measurable division of labour. It is a place where authors can slow down, reconstruct the workflow, and become more articulate about a collaboration that fluency has made too easy to forget.

The label is a small instrument with a large aspiration: that pausing to describe one's workflow might become, over time, a habit through which a scholar learns to inhabit their own prosthetic condition with something approaching honesty. It will not resolve distributed authorship into clean proportions. It will not prevent dishonesty or guarantee accurate memory. Its contribution lies elsewhere. Each act of classification returns to the surface a question that fluent composition has already begun to absorb: how did this work come to be, and through what relations of human and machine labour did it find expression?

Acknowledgements

The author owes special thanks to David Leach, whose offhand provocation at the UVic Libraries “AI in Research” forum — that research might need something like a label showing how much AI was in it — provided the generative spark for the AI Usage Label maker tool. His later correspondence and encouragement helped sustain the project as it moved from a speculative idea into a working disclosure instrument. The author is deeply grateful to Ray Siemens at the University of Victoria’s Electronic Textual Cultures Lab for his generous feedback, intellectual support, and invaluable suggestions for the tool’s development, including refinements to its staged scale and attention to questions of reproducibility. Thanks are also due to Matt Huculak, whose encouragement helped bring the project to a wider scholarly and library audience, and to Rich McCue, whose public use of the tool with students and colleagues offered

an important early demonstration of its practical value. Their collective insights strengthened the project from its initial provocation through its design, circulation, and theoretical articulation.

References

Ahmed, S. (2004). *The cultural politics of emotion*. Edinburgh University Press.

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, 13(3), 219–235.

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>

Chan, C. K. Y. (2024). Students' perceptions of "AI-giarism": Investigating changes in understandings of academic misconduct. *Education and Information Technologies*, 30, 8087–8108. <https://doi.org/10.1007/s10639-024-13151-7>

Chaudoir, S. R., & Fisher, J. D. (2010). The disclosure processes model. *Psychological Bulletin*, 136(2), 236–256. <https://doi.org/10.1037/a0018193>

Clark, A. (2003). *Natural-born cyborgs: Minds, technologies, and the future of human intelligence*. Oxford University Press.

Clark, A., & Chalmers, D. J. (1998). The extended mind. *Analysis*, 58(1), 7–19.
<https://doi.org/10.1093/analys/58.1.7>

Goffman, E. (1963). *Stigma: Notes on the management of spoiled identity*. Simon & Schuster.

Haraway, D. (1991). A cyborg manifesto: Science, technology, and socialist-feminism in the late twentieth century. In *Simians, cyborgs, and women: The reinvention of nature* (pp. 149–181). Routledge.

Hutchins, E. (1995). *Cognition in the wild*. MIT Press.

Ihde, D. (1990). *Technology and the lifeworld: From garden to earth*. Indiana University Press.

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, 114(1), 3–28.

Link, B. G., & Phelan, J. C. (2001). Conceptualizing stigma. *Annual Review of Sociology*, 27, 363–385. <https://doi.org/10.1146/annurev.soc.27.1.363>

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–16. <https://doi.org/10.1145/3313831.3376727>

Lucas, G. M., Becerik-Gerber, B., & Roll, S. C. (2024). Calibrating workers' trust in intelligent automated systems. *Patterns*, 5(9), 101045. <https://doi.org/10.1016/j.patter.2024.101045>

Lund, B., Mannuru, N. R., Teel, Z. A., Lee, T. H., Ortega, N. J., Simmons, S., & Ward, E. (2025). Student perceptions of AI-assisted writing and academic integrity. *AI in Education*, 1(1), 2. <https://doi.org/10.3390/aieduc1010002>

Mollick, E. R., & Mollick, L. (2023). *Assigning AI: Seven approaches for students, with prompts* [Wharton School Research Paper]. <https://doi.org/10.2139/ssrn.4475995>

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy. *Computers and Education: Artificial Intelligence*, 2, 100041. <https://doi.org/10.1016/j.caeai.2021.100041>

ORCID. (2022). ORCID's first decade: From startup to sustainability. <https://info.orcid.org/orcid-first-decade-history/>

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation. *Human Factors*, 52(3), 381–410. <https://doi.org/10.1177/0018720810376055>

U.S. Food and Drug Administration. (2020). History of nutrition labeling. <https://www.fda.gov/food/nutrition-facts-label/history-nutrition-facts-label>

U.S. Food and Drug Administration. (2016). Final rule: Changes to the nutrition facts label (21 CFR Parts 101 and 130). <https://www.federalregister.gov/documents/2016/05/27/2016-11867>

Verbeek, P. P. (2005). *What things do: Philosophical reflections on technology, agency, and design*. Penn State University Press.