

The Cognitive Prosthetic: Why AI Usage Disclosure Matters

A Philosophical Framework for the AI Usage Facts Label

Faraz Forghan Parast

Electronic Textual Cultures Lab, University of Victoria, 2025

Work in Progress

Abstract

Large language models have become embedded in scholarly workflows at a pace that outstrips institutional capacity to understand what has changed. This paper argues that generative AI functions as a cognitive and linguistic prosthetic, one that transforms the phenomenology of authorship, the distribution of epistemic labor, and the affective atmosphere surrounding scholarly production. Drawing on philosophy of technology, embodied cognition, affect theory, disability studies, and empirical research on human-AI collaboration, the paper develops a framework for understanding why transparent disclosure of AI involvement is an epistemological and pedagogical necessity. The AI Usage Facts label—a customizable tool for documenting AI contribution across research stages—emerges from this framework as both reflective practice and epistemic infrastructure. The paper demonstrates how structured disclosure addresses source-monitoring failures, calibrates reader trust, reduces stigma through granularity, and cultivates the literacy required for responsible use of cognitive prosthetics in scholarship.

Introduction: The Provenance Problem

Scholarship has always depended on a tacit contract: the reader assumes that the text before them emerged from a process legible enough to warrant trust. Citations signal debts and sources. Methods sections describe procedures. Acknowledgments name collaborators and funders. These conventions exist because provenance matters. Knowing how a claim came to be shapes how we evaluate it, how we teach it, and whether we build on it. Reproducibility, attribution, and verification form the cornerstones of this contract. When a reader encounters a claim, they assume not only that the author can account for its origins but that another scholar, following the same methods and sources, could arrive at a comparable result. This assumption underwrites the cumulative nature of knowledge production.

Generative AI disrupts this contract at scale. A manuscript produced with substantial assistance from a large language model can be indistinguishable on its surface from one composed entirely by a human author. The prose may be fluent, the structure coherent, the references plausible. Yet the conditions under which that text came into being—the distribution of cognitive labor, the points at which verification occurred, the risks of hallucination or stylistic homogenization—remain invisible. The reader lacks the

information necessary to calibrate trust. The author may lack a stable account of their own contribution. The institution lacks a vocabulary for teaching, evaluating, or even discussing what has happened. Most critically, the black-box nature of these systems means that results are often non-reproducible. Two scholars using the same prompt with the same model at different times may receive substantively different outputs. The stochastic sampling mechanisms that generate text introduce variability that cannot be fully controlled or documented (Bender et al., 2021). For disciplines grounded in reproducibility as a methodological cornerstone—particularly the hard sciences and solution-oriented fields—this opacity represents not merely an ethical concern but a foundational threat to the validity of knowledge claims.

This paper argues that the problem is deeper than ethics or policy. Generative AI constitutes a cognitive and linguistic *prosthetic*, one that mediates thought and language in ways that reshape the user's relationship to their own agency, memory, and voice. Like all prosthetics, AI does more than augment or replace a function. It alters the user's body schema, their perceptual field, and the affective atmosphere in which they work (Sobchack, 2006; Murray, 2004). The text that emerges from human-AI collaboration exists in a new ontological category: a hybrid form that current scholarly infrastructure cannot adequately describe.

The AI Usage Facts label, available at <https://ailabel.netlify.app>, is a response to this gap. Modeled on the nutrition facts label, it provides a standardized format for voluntarily disclosing AI involvement across the stages of research and writing. Users specify which tasks involved AI assistance, to what degree, under what quality controls, and with what tools. The label generates a percentage index summarizing total AI contribution, exports machine-readable metadata, and includes warnings about residual risks. The tool also features a customizable workflow editor, allowing disciplines to define their own stage models and epistemic weights.

This paper develops the philosophical and pedagogical rationale for such a tool. It treats disclosure as a practice that serves three overlapping functions: it provides readers with epistemic context they currently lack; it cultivates in authors the reflective capacity required to use cognitive prosthetics responsibly; and it intervenes in the affective field surrounding AI-assisted scholarship, replacing binary stigma with structured description. The argument proceeds in four movements. First, it establishes AI as a cognitive prosthetic and explores what that framing reveals about authorship, embodiment, and mediation. Second, it examines the literacy gap: the absence of shared practices, vocabularies, and norms for navigating prosthetic scholarship. Third, it analyzes the affective and cognitive dynamics that make undisclosed AI use epistemically and socially corrosive. Fourth, it presents the AI Usage Facts label as a practical instantiation of the framework, demonstrating how design choices encode philosophical commitments about transparency, granularity, and disciplinary plurality.

The Prosthetic Condition: AI and the Reconfiguration of Authorship

Prosthetics are typically understood through a deficit model: something is absent, and technology compensates. A prosthetic limb replaces a biological one. A hearing aid restores diminished auditory capacity. This framing treats the prosthetic as a neutral substitute, aiming to return the user to a prior state of normalcy. Disability studies and phenomenology have long contested this view. Prosthetics do more than fill gaps: they create new configurations of embodiment, new modes of perception, and new relationships between the user and their environment (Murray, 2004; Sobchack, 2006; Kurzman, 2001). The runner with carbon-fiber blades does not simply regain the ability to run; they run differently, with a different gait, different biomechanics, and a different felt sense of motion. The cochlear implant user does not hear as a hearing person hears; they experience sound through a device that translates vibration into electrical signals, producing a qualitatively distinct auditory world.

Merleau-Ponty's (1945/2012) classic example of the blind person's cane illustrates the phenomenological stakes. The cane is experienced as an extension of the body's perceptual apparatus. The user does not feel the cane in their hand and then infer what lies at its tip; they feel the world through the cane, as if their sense of touch had migrated to its end. The cane becomes incorporated into the body schema, the pre-reflective sense of where one's body begins and ends. This incorporation is what allows the prosthetic to function transparently: it recedes from conscious attention and becomes part of the perceptual background.

Large language models function as prosthetics for cognition and language. When a scholar uses Claude, GPT, or Gemini to draft a paragraph, synthesize literature, or generate research questions, they are extending their cognitive reach into domains that may exceed their unaided capacity. The model provides access to linguistic patterns drawn from billions of documents. It offers syntactic fluency, stylistic variation, and associative leaps that the user might achieve slowly or with difficulty. It distributes the labor of composition across a human-machine assemblage, where the boundary between "my thought" and "its output" becomes permeable and often invisible. This distribution of cognitive labor, or what Hutchins (1995) termed "distributed cognition," is where cognitive processes are spread across individuals, artifacts, and time. Recent empirical work confirms that users flexibly distribute cognitive demands between internal and external resources when solving problems (Grinschgl & Neubauer, 2022), and that AI tools enable forms of cognitive offloading that reshape how users approach tasks (Singh et al., 2025).

This permeability is what makes AI a prosthetic rather than a tool in the conventional sense. A hammer remains external to the user's body schema; it is wielded, set down, and picked up again. A prosthetic limb or a cochlear implant becomes part of how the user

inhabits the world. Similarly, scholars who write regularly with AI assistance begin to think through the model. They internalize its patterns, adopt its phrasings, and rely on its associative logic. The model's strengths become their strengths; its limitations become their limitations. The text that emerges is neither purely human nor purely machine-generated. It is a product of prosthetic authorship, where agency is distributed and the locus of origination is genuinely unclear. Clark (2003) describes this as the "extended mind," where cognitive processes extend beyond the biological brain to incorporate external tools and technologies.

This distribution has consequences that extend beyond the moment of composition. Prosthetics alter what users can do and how they perceive themselves and how others perceive them. A prosthetic limb changes the wearer's gait, their balance, and their sense of bodily integrity. It also changes how they are seen: as disabled, as augmented, as cyborg, as inspirational, as pitiable, depending on the social and cultural frame. The prosthetic becomes a site where identity, capability, and social meaning converge (Sobchack, 2006; Haraway, 1991; Reeve, 2012).

AI-assisted authorship operates similarly. The scholar who uses AI extensively may experience shifts in their sense of authorial voice, their confidence in their own claims, and their memory of what they wrote versus what the model suggested. They may feel relief at the speed and fluency AI provides, or anxiety about whether the work is "really theirs." They may forget which ideas originated in their own thinking and which emerged from iterative prompting. These shifts are phenomenological: they concern the lived experience of being an author under conditions of distributed cognition.

At the same time, the social meaning of AI-assisted work is contested and unstable. Recent empirical studies reveal that students and faculty hold complex and often contradictory attitudes toward AI use in academic writing. Lund et al. (2025) found that students who believe AI-generated work constitutes cheating are substantially less likely to use it, while policy awareness had no significant effect on behavior. Chan (2024) documented that students clearly disapprove of direct AI content generation but hold ambivalent attitudes toward subtler uses, a phenomenon she terms "AI-giarism." In some contexts, AI use signals efficiency and technological fluency. In others, it signals laziness, incompetence, or fraud. The same text can be read as innovative or as "slop," depending on whether the reader knows AI was involved and how they interpret that involvement. This interpretive instability creates an affective atmosphere of suspicion, defensiveness, and stigma that shapes how scholars approach AI and how they narrate their own practices.

Haraway's (1991) cyborg offers a useful conceptual resource here. The cyborg refuses the boundary between human and machine, treating hybridity as a condition. Scholars using AI are already cyborgs in this sense. The question is whether they acknowledge that

condition, hide it, or deny it. Acknowledgment opens the possibility of developing practices, norms, and literacies appropriate to prosthetic scholarship. Denial forecloses that possibility and perpetuates a fiction of autonomous authorship that no longer describes how texts are actually produced.

The prosthetic framing thus reveals something crucial: AI-assisted authorship is a transformation of embodiment, perception, and social identity, mediated by technology that becomes incorporated into the user's cognitive and linguistic apparatus. Disclosure becomes necessary because the surface artifact—the text itself—no longer carries sufficient information about the conditions of its production. The reader cannot tell from the prose alone whether it emerged from solitary human effort, iterative human-AI collaboration, or wholesale generation with minimal human oversight. The author may struggle to reconstruct that process themselves, especially as time passes and memory fades. Transparent disclosure is the mechanism by which prosthetic authorship becomes legible, both to others and to the author themselves.

The Literacy Gap: Rehabilitation for a Cognitive Prosthetic

Every prosthetic requires a period of adjustment, training, and skill acquisition. A person fitted with a prosthetic limb does not immediately walk with ease. There is disorientation as the body recalibrates its sense of balance and proprioception. There is deliberate practice as new movement patterns are learned and old habits unlearned. There is gradual integration as the prosthetic becomes incorporated into the body schema, shifting from a consciously managed attachment to a transparent extension of the self. This process is supported by physical therapy, occupational therapy, and peer communities who share knowledge about living with prosthetics. The rehabilitation is recognized as essential: the technology alone does not confer competence.

Generative AI has been deployed as a cognitive prosthetic with essentially no equivalent rehabilitation infrastructure. Billions of people now have access to large language models capable of drafting essays, generating code, summarizing research, and answering complex questions. These tools mediate thought and language in ways that are intimate and consequential. Yet there is no widespread curriculum for learning to use them well, no shared vocabulary for describing degrees or types of AI involvement, no established practices for verifying outputs or reflecting on how AI shapes thinking. Users are left to develop their own ad hoc strategies, often learning through intuition and trial and error what the models do well, where they fail, and how to integrate their outputs into human workflows. Stanford (2025) found that students using AI tools for academic writing improved lower-order skills like grammar and fluency, but struggled with higher-order skills like critical thinking and originality, suggesting that AI support can create skill atrophy when used without proper guidance.

This absence of structured literacy creates predictable problems. Users plateau at basic competence: they learn to write prompts that elicit useful responses, but they do not develop the evaluative skills needed to distinguish high-quality outputs from plausible-sounding errors. They do not cultivate the metacognitive awareness required to notice when AI is subtly reshaping their thinking, narrowing their vocabulary, or homogenizing their style. They do not acquire the reflective capacity to articulate what AI contributed to a given piece of work, because they have no framework for tracking or categorizing that contribution. The result is a population of prosthetic users who can operate the technology but who lack the deeper literacies that would allow them to use it responsibly, critically, and transparently.

Effective use of AI as a cognitive prosthetic involves multiple layers of competence. At the most basic level, users must learn prompt craft: how to elicit useful outputs through clear, specific, and well-structured queries. This skill is widely discussed and relatively easy to acquire. The next layer involves output evaluation: the ability to assess whether a given AI-generated text is accurate, relevant, and appropriate for the intended purpose. This requires domain knowledge, critical reading skills, and familiarity with the model's tendencies toward hallucination, bias, and stylistic homogenization. The third layer is integration judgment: knowing what to keep, what to modify, and what to discard from AI outputs, and how to weave those elements into a coherent whole that reflects the author's voice and intentions. The fourth layer is self-awareness: recognizing how AI is shaping one's own thinking, what cognitive labor is being offloaded, and what risks or limitations accompany that offloading. The fifth and deepest layer is disclosure capacity: being able to articulate AI's role accurately and meaningfully, both to oneself and to others.

Most users operate at the first or second layer. The AI Usage Facts label is designed to scaffold movement toward the fourth and fifth layers. By requiring users to reconstruct their workflow, classify each stage by degree of AI involvement, and specify quality controls, the label functions as a reflective practice. It asks users to slow down and remember: Which tasks involved AI? To what extent? Under what conditions? With what verification? This process of structured recall produces self-knowledge. It makes visible patterns that might otherwise remain tacit or forgotten. It creates a vocabulary for describing human-AI collaboration that can be shared, taught, and refined over time.

The workflow editor within the AI Usage Facts label maker extends this pedagogical function. The tool provides default stage models based on established frameworks like the CRediT taxonomy, which breaks research into stages such as conceptualization, methodology, investigation, analysis, writing, and review. Users can adopt these defaults or customize them to fit their discipline, methodology, or project type. A creative writer might define stages as ideation, drafting, revision, and editing. An empirical scientist might use literature review, hypothesis formation, data collection, statistical analysis,

interpretation, and manuscript preparation. A historian might include archival research, source analysis, narrative construction, and argumentation. The flexibility matters because different disciplines have different epistemic risk profiles. AI involvement in data analysis carries different implications than AI involvement in brainstorming keywords for a literature search. A customizable workflow allows communities to articulate those differences and to develop norms appropriate to their practices.

This customization also makes the label a teachable instrument. Faculty can use it to guide students through the process of documenting their research practices. Librarians can incorporate it into information literacy instruction, helping researchers understand how to evaluate and disclose the provenance of their work. Writing centers can use it as a reflective tool, prompting writers to think critically about where AI assistance is helpful and where it might undermine learning or originality. The label becomes infrastructure for developing the literacies that prosthetic scholarship requires.

The rehabilitation metaphor thus captures something essential about the current moment. Scholars have been handed a powerful cognitive prosthetic and told to figure it out. The AI Usage Facts label offers a structured, deliberate practice for that figuring-out: a way to develop competence, cultivate awareness, and build shared norms through repeated, reflective engagement with the question of what AI contributed and how.

Affect, Bias, and the Epistemology of Undisclosed AI Use

The case for disclosure rests on more than ethics or pedagogy. It also addresses a set of cognitive and affective dynamics that make undisclosed AI use epistemically corrosive. These dynamics operate at multiple levels: in the author's relationship to their own work, in the reader's interpretation of that work, and in the broader social atmosphere surrounding AI-assisted scholarship.

One of the most troubling findings from cognitive psychology concerns source monitoring: the ability to remember where a piece of information came from. Humans are remarkably poor at this task, especially when the information is presented in fluent, contextually appropriate language (Johnson et al., 1993). We misattribute ideas we read to our own thinking. We forget whether a claim originated in a conversation, a document, or our own reasoning. This phenomenon is amplified with AI because the outputs are designed to be seamless. A large language model generates text that mirrors the user's style, fits the context, and flows naturally into the surrounding prose. Over time, and especially after iterative revision, the boundary between "what I wrote" and "what the model suggested" becomes genuinely unclear. The author may sincerely believe they composed a passage that was in fact generated by AI, or they may forget that a key idea emerged from a prompt rather than from their own reading or reflection.

This source-monitoring failure creates an epistemic asymmetry between author and reader. The surface artifact—the text on the page—is identical whether the reader knows about AI involvement or remains ignorant of it. Yet the interpretive situation is radically different. The author, even if their memory is imperfect, has lived through the process of generation, selection, and revision. They carry implicit knowledge of where uncertainty exists, which claims were verified, and which passages were accepted wholesale from AI outputs. The reader has access to none of this. They see only the finished text and attribute coherence, expertise, and intentionality to the human author. They form judgments about the author's knowledge, competence, and credibility based on assumptions that may be false. This asymmetry constitutes what is essentially an epistemic injustice by omission: the reader lacks information necessary for appropriate evaluation, and the author possesses information they do not share.

The problem is compounded by the fluency heuristic: the tendency to judge fluent, well-structured prose as more true, more credible, and more authoritative than awkward or hesitant language (Alter & Oppenheimer, 2009). Large language models are optimized for fluency. They produce smooth, confident, grammatically polished text even when the underlying claims are speculative, unsupported, or hallucinated. This creates a dangerous mismatch: the surface style signals expertise and certainty, while the actual epistemic status of the claims may be far weaker. Both authors and readers are vulnerable to this mismatch. Authors may trust AI-generated text because it sounds authoritative. Readers may trust the author because the prose is polished. The fluency trap operates at both ends of the communication chain, systematically inflating confidence beyond what the evidence warrants.

A related problem emerges from research on human-automation interaction. When automated systems perform well most of the time, human operators become poor at detecting failures (Parasuraman & Manzey, 2010). This phenomenon, known as automation complacency or automation bias, occurs because vigilance is cognitively costly. If 95% of AI suggestions are useful, users stop scrutinizing the remaining 5%. Verification effort declines as trust increases. The errors that do occur become harder to catch because the user has outsourced the critical reading that would detect them. Romeo and Conti (2025) found that automation bias is particularly pronounced when users encounter system strengths before weaknesses, leading to increased reliance and more errors. In the context of AI-assisted writing, this means that authors may genuinely lack knowledge of whether a particular claim is accurate, whether a citation is real, or whether a line of reasoning is sound. They have accepted the output because it seemed plausible and because checking every detail is exhausting. Recent work on trust calibration in AI systems emphasizes that appropriate reliance requires users to distinguish cases where they can trust AI predictions from those where they should remain skeptical (Turner et al., 2025; Lucas et al., 2024).

These cognitive dynamics traverse affective ones. Affect, in the sense developed by Massumi (2002) and Ahmed (2004), is the pre-personal intensity that moves bodies into action and the way certain feelings come to "stick" to objects, signs, and practices. In contemporary discourse, "AI-assisted" has become sticky with a cluster of negative affects: suspicion, contempt, anxiety about low-effort work, fear of hallucination, and moral panic about theft and inauthenticity. Meanwhile, "human-authored" remains sticky with authenticity, effort, and deservingness. This affective distribution shapes how scholars approach AI and how they narrate their own practices. Many users feel shame or impostor syndrome when they use AI extensively, even when that use is appropriate and well-controlled. Others feel defensive, anticipating judgment or accusations of cheating. Still others adopt a posture of bravado, insisting that AI is "just a tool" and that concerns about disclosure are overblown.

These affective responses are intensified by the binary framing that dominates public discourse. A piece of work is either "human" or "AI-generated," with little room for the gradient of involvement that actually characterizes most AI-assisted scholarship. This binary collapses nuance and produces a climate of suspicion where any AI involvement is treated as contamination. The result is a chilling effect: scholars who use AI responsibly may hide that use to avoid stigma, while those who use it irresponsibly may do the same to avoid consequences. The lack of a shared vocabulary for describing degrees and types of AI involvement makes it difficult to have productive conversations about what constitutes appropriate use.

Ahmed's (2004) concept of affective economies is useful here. Affects circulate through social fields, sticking to some objects and sliding off others, shaping what feels possible, desirable, or shameful. The current affective economy around AI-assisted scholarship is characterized by fear, defensiveness, and moral polarization. Disclosure practices can intervene in this economy by redistributing what sticks where. A structured, granular disclosure format like the AI Usage Facts label breaks the binary. It replaces "I used AI" (which sticks to shame and suspicion) with "AI contributed this much to this manuscript, primarily in literature review and editing, under the following quality controls" (which sticks to transparency, competence, and care). The percentage index, the stage breakdown, and the quality controls together create a different affective orientation: one where AI use is described rather than confessed or hidden.

This reorientation has epistemic consequences. When authors can describe their AI use in structured terms, they are more likely to disclose it. When readers receive that disclosure in a standardized format, they can calibrate their trust appropriately. The dynamic can be seen easily: an author who has used AI modestly—perhaps for brainstorming keywords, polishing prose, or checking grammar—may feel anxious about whether their work "counts" as their own. Seeing that AI contributed only 12% to the final manuscript, and that all AI outputs were verified and integrated under human oversight, provides

reassurance. It restores a sense of proportion and agency. For the reader, that same 12% figure provides context. It signals that the work is substantially human-authored, that AI played a supporting role, and that the author has been transparent about that role. The granularity matters because it resists the flattening logic of the binary: any AI involvement equals total AI generation equals fraud.

The affective and cognitive dynamics thus reinforce each other. Source-monitoring failures make it hard for authors to know what they contributed. Fluency heuristics make it hard for readers to detect when claims are unsupported. Automation complacency makes it hard for both to maintain vigilance. Stigma makes it hard to talk about any of this openly. The AI Usage Facts label is an attempt to address all of these problems simultaneously. It scaffolds source monitoring by requiring retrospective reconstruction of the workflow. It counters fluency heuristics by making visible the conditions under which fluent text was produced. It mitigates automation complacency by prompting users to specify quality controls. It intervenes in the affective economy by providing a vocabulary for disclosure that resists stigma. The label is thus both an epistemic tool and an affective one: it changes what can be *known* and what can be *felt* about AI-assisted scholarship.

The AI Usage Facts Label: Design as Philosophy

The AI Usage Facts label translates the conceptual framework developed above into a practical instrument. Its design choices encode philosophical commitments about transparency, granularity, disciplinary plurality, and the relationship between disclosure and self-knowledge. This section examines those choices and demonstrates how the tool operationalizes the arguments for prosthetic literacy, affective reorientation, and epistemic infrastructure.

The nutrition facts metaphor is foundational. Nutrition labels emerged in response to a similar problem: consumers lacked information about what was in their food, and that lack of information made informed choice impossible. The label standardizes disclosure in a recognizable format, quantifies components that were previously vague, itemizes ingredients, warns of potential risks, and empowers consumers to make decisions aligned with their values and needs. The metaphor also carries a message: processed food is widespread and serves legitimate purposes, but people deserve to know what they are consuming. The label informs rather than moralizes.

The AI Usage Facts label applies this logic to scholarly texts. It standardizes disclosure of AI involvement across research stages, quantifies contribution through a percentage index, itemizes the tools and models used, warns of residual risks like hallucination or bias, and empowers readers to evaluate the work appropriately. The label makes AI use legible so that others can form their own judgments. This stance is deliberate. The goal is

to build infrastructure for informed evaluation, reducing both unwarranted suspicion and unwarranted trust.

The percentage index is perhaps the most visible and most contested feature of the label. Critics might object that authorship cannot be quantified, that the boundary between human and AI contribution is too blurred to measure, or that any such metric will be gamed or misunderstood. These objections have merit, but they miss the function the percentage serves. The index is a communicative compression of a disclosed workflow. It summarizes, in a single number, the weighted average of AI involvement across all stages of research and writing. That number is derived from user self-report: for each task within each stage, the user indicates whether AI was involved (yes, partial, or no; or on a five-degree scale of 0%, 25%, 50%, 75% and 100% AI participation), and the tool calculates stage-level percentages based on those responses. Stages are then weighted according to their epistemic significance, with data analysis and writing typically receiving higher weights than ideation or formatting.

The percentage is an index, a summary statistic that points back to the underlying workflow disclosure. It functions like a confidence interval or a p-value: a compressed representation of a more complex reality, useful for quick interpretation but always requiring context. The value of the index lies in its ability to resist binary thinking. An 8% AI contribution is meaningfully different from an 80% contribution, and both are different from 0% or 100%. The granularity allows for proportional responses. A reader encountering an 8% label might proceed with confidence, knowing that AI played a minor role. A reader encountering an 80% label might read more critically, checking claims and citations with greater care. Both responses are appropriate; the label provides the information needed to calibrate them.

The workflow editor is the mechanism that makes this system flexible and pedagogically generative. The tool provides default stage models based on established taxonomies like CRediT, which divides research into conceptualization, methodology, investigation, formal analysis, writing, and review. Each stage contains tasks that can be marked as involving AI fully, partially, or minimally. Users can adopt these defaults or create custom workflows tailored to their discipline or project. A creative writer might define stages as ideation, drafting, revision, and editing, with tasks like character development, dialogue writing, and line editing. An empirical scientist might use literature review, hypothesis formation, experimental design, data collection, statistical analysis, interpretation, and manuscript preparation. A historian might include archival research, source analysis, contextualization, narrative construction, and argumentation.

This customizability serves two purposes. First, it acknowledges that different kinds of scholarship have different structures and different points where AI mediation matters. A one-size-fits-all disclosure form would either be too generic to be useful or too specific to

fit diverse practices. The workflow editor allows communities to negotiate what counts as a meaningful stage, what tasks belong to each stage, and how those stages should be weighted. This negotiation is itself a form of norm-building: it requires explicit discussion of what matters epistemically in a given domain. Second, the workflow editor makes the label a pedagogical tool. Faculty can use it to teach students how to break down a research project into stages, how to think critically about where AI might be helpful or harmful, and how to document their practices transparently. The act of filling out the label becomes a reflective exercise, prompting users to reconstruct their workflow and to notice patterns they might otherwise overlook.

The quality controls section of the label addresses the verification problem. AI outputs are probabilistic and prone to errors: hallucinated citations, plausible-sounding falsehoods, stylistic homogenization, and unintentional plagiarism. The label prompts users to indicate which quality controls they have implemented: human verification of AI-generated claims, plagiarism scanning, citation cross-checking, assessment of data leakage risk, and ethics or IRB review where applicable. These controls are presented as checkboxes, making them easy to report and easy for readers to scan. The presence or absence of these controls provides crucial context. A manuscript with 60% AI contribution and rigorous verification is epistemically different from one with 60% AI contribution and minimal oversight. The label makes that difference visible.

The warnings section serves a similar function. The label includes standardized warnings about risks associated with AI use: outputs may contain hallucinated claims if unverified, models may omit literature published after their training cutoff, image generation tools may encode stylistic biases, and so on. These warnings are displayed prominently on the label itself and included in the machine-readable JSON export. They function as a reminder that disclosure is a starting point for evaluation, a signal that the author has thought about these risks and that the reader should remain appropriately critical.

The export options reflect the label's dual function as human-readable communication and machine-readable metadata. Users can export the label as a PNG image suitable for inclusion in supplementary materials or appendices. They can also export it as JSON, a structured data format that can be ingested by repositories, indexing services, or future verification tools. The machine-readable format opens possibilities for large-scale analysis of AI use patterns, for automated checks of disclosure completeness, and for integration with scholarly infrastructure like ORCID, DOI systems, or institutional repositories. The label is designed to be both immediately useful and forward-compatible with emerging norms and technologies.

The design philosophy can be summarized as follows: the label treats disclosure as a practice that produces value through repetition, reflection, and standardization. It embraces approximate self-report as better than silence. It describes rather than judges. It

allows customization rather than imposing uniformity. It provides granularity rather than binary categories. It generates machine-readable metadata rather than remaining opaque. Each of these design choices follows from the conceptual framework developed earlier: AI as prosthetic, disclosure as literacy, and transparency as infrastructure for trust.

Conclusion: Disclosure as Epistemic Virtue

The argument developed in this paper can be stated simply: generative AI functions as a cognitive and linguistic prosthetic, one that transforms the conditions under which scholarly texts are produced. That transformation is phenomenological, epistemic, and affective. It changes how authors experience their own agency, how readers interpret texts, and how communities understand what counts as legitimate scholarship.

Transparent disclosure of AI involvement is necessary because the surface artifact no longer carries sufficient information about its own provenance. The AI Usage Facts label provides a practical mechanism for such disclosure, one that is standardized, customizable, and designed to cultivate the reflective literacies that prosthetic scholarship requires.

The case for disclosure rests on three overlapping arguments. The first is epistemic: readers need context to calibrate trust, and authors need practices that scaffold accurate self-knowledge. The second is pedagogical: institutions need tools for teaching responsible AI use, and communities need shared vocabularies for discussing what that responsibility entails. The third is affective: the current climate of stigma and suspicion around AI use is corrosive, and structured disclosure can intervene in that climate by replacing binary judgment with proportional description.

The label is a small intervention in a large and rapidly evolving landscape. It cannot solve all the challenges posed by AI-assisted scholarship. It cannot prevent dishonesty, guarantee accuracy, or eliminate the interpretive difficulties that arise when authorship is distributed across human and machine agents. What it can do is create a space for reflection, a vocabulary for description, and a norm of transparency that makes prosthetic authorship legible and discussable. These are modest goals, but they are foundational. Scholarship depends on trust, and trust depends on legibility. The AI Usage Facts label is an instrument for restoring that legibility under conditions where it has been disrupted.

The tool also embodies a particular stance toward technological change: one that treats AI as a prosthetic that reconfigures how we think, write, and relate to our own work. Like all prosthetics, it requires rehabilitation, literacy, and ongoing adjustment. The label is part of that rehabilitation. It asks users to slow down, to remember, to classify, and to disclose. This friction is deliberate. It creates space for the self-awareness that responsible prosthetic use demands.

Looking forward, the label's machine-readable format and customizable workflow editor suggest possibilities for broader integration into scholarly infrastructure. Journals could require AI Usage Facts labels as part of submission processes. Repositories could index them as metadata. Institutions could use them to track AI adoption patterns and to identify where additional training or support is needed. Disciplines could develop their own stage models and epistemic weights, tailoring the disclosure framework to their specific practices and values. The label is designed to be extensible, adaptable, and capable of evolving alongside the technologies and norms it seeks to document.

The ultimate goal is the cultivation of epistemic virtues appropriate to an era of distributed cognition. Intellectual humility: recognizing the limits of one's own knowledge and the contributions of non-human agents. Intellectual responsibility: verifying claims, disclosing sources, and maintaining vigilance over automated processes. Intellectual honesty: representing one's work accurately, resisting the temptation to claim credit for labor one did not perform. These virtues have always mattered in scholarship. What has changed is the context in which they must be practiced. The AI Usage Facts label helps practice them under new conditions, where the boundary between human and machine contribution is genuinely unclear and where transparency requires structured effort.

Scholars using AI are already cyborgs, already prosthetic users, navigating a hybrid mode of authorship that existing norms and infrastructures struggle to accommodate. The question is whether that condition will be acknowledged, hidden, or denied. Acknowledgment is the path toward developing practices, literacies, and norms that fit the reality of how texts are now produced. The AI Usage Facts label is an invitation to that acknowledgment: to slow down, to reflect, to disclose, and to join a community building the infrastructure for prosthetic scholarship. The work is just beginning, but it begins with honesty about what we know, how we know it, and where our thinking has been shaped by the tools we use.

Acknowledgements

I am grateful to Ray Siemens at the Electronic Textual Cultures Lab at the University of Victoria for his generous feedback and invaluable suggestions for the AI Usage Facts tool. David Leach's initial provocation at the "AI in Research" forum at UVic Libraries and later correspondence proved instrumental in shaping the project's direction. Matt Huculak and Rick McCue both kindly saw potential in this work early on and have provided crucial support throughout its ongoing development. Their collective insights have strengthened both this conceptual framework and its practical instantiation.

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.
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. <https://doi.org/10.1145/3442188.3445922>
- 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>
- Clark, A. (2003). *Natural-born cyborgs: Minds, technologies, and the future of human intelligence*. Oxford University Press.
- Grinschgl, S., & Neubauer, A. C. (2022). Supporting cognition with modern technology: Distributed cognition today and in an AI-enhanced future. *Frontiers in Artificial Intelligence*, 5, 908261. <https://doi.org/10.3389/frai.2022.908261>
- 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.
- Kurzman, S. L. (2001). Presence and prosthesis: A response to Nelson and Wright. *Cultural Anthropology*, 16(3), 374-387.
- 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: Ethical concerns, academic misconduct, and use of generative AI in higher education. *AI in Education*, 1(1), 2. <https://doi.org/10.3390/aieduc1010002>
- Massumi, B. (2002). *Parables for the virtual: Movement, affect, sensation*. Duke University Press.
- Merleau-Ponty, M. (2012). *Phenomenology of perception* (D. A. Landes, Trans.). Routledge. (Original work published 1945)
- Murray, C. D. (2004). An interpretative phenomenological analysis of the embodiment of artificial limbs. *Disability and Rehabilitation*, 26(16), 963-973.
- Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3), 381-410.
- Reeve, D. (2012). Cyborgs, cripples and iCrip: Reflections on the contribution of Haraway to disability studies. In D. Goodley, B. Hughes, & L. Davis (Eds.), *Disability and social theory: New developments and directions* (pp. 91-111). Palgrave Macmillan.

Romeo, G., & Conti, D. (2025). Exploring automation bias in human-AI collaboration: A review and implications for explainable AI. *AI & Society*, 41, 259-278. <https://doi.org/10.1007/s00146-025-02422-7>

Singh, A., Taneja, K., Guan, Z., & Ghosh, A. (2025). Protecting human cognition in the age of AI. arXiv preprint arXiv:2502.12447.

Sobchack, V. (2006). A leg to stand on: Prosthetics, metaphor, and materiality. In M. Smith & J. Morra (Eds.), *The prosthetic impulse: From a posthuman present to a biocultural future* (pp. 17-41). MIT Press.

Stanford, J. (2025). How disruptive is AI in students' academic writing in higher education? *Studies in Technology Enhanced Learning*, 4(1). <https://doi.org/10.21428/8c225f6e.b1d05ed3>

Turner, A., Kaushik, M., Huang, M., & Varanasi, S. (2025). Calibrating reliance on automated advice: Transparency and trust calibration feedback. *International Journal of Human-Computer Interaction*, 41(23), 14723-14733. <https://doi.org/10.1080/10447318.2025.2487861>

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