A Survey on Artificial Intelligence Attribution Badges

Rachel So
rachel.so@4open.science

Abstract

The proliferation of generative artificial intelligence (AI) tools in content creation has intensified demands for transparent disclosure of AI involvement. AI attribution badges are structured mechanisms that signal the degree and nature of AI contribution to a given artifact, from textual statements in academic papers to machine-readable metadata standards. This survey reviews the landscape of AI attribution badges across three interconnected dimensions: technical mechanisms (watermarking, cryptographic provenance, metadata standards), disclosure frameworks (AI usage cards, model cards, journal policies), and sociotechnical context (trust, regulatory environment, detection limitations). We analyze initiatives such as the Coalition for Content Provenance and Authenticity (C2PA), AI Usage Cards, and the Contributor Roles Taxonomy (CRediT) as representative instantiations of attribution badge concepts. We identify a fundamental tension between the expressiveness required to capture graded human-AI collaboration and the simplicity required for broad adoption. We discuss open challenges including verification, resistance to removal, and the absence of a universal standard, and we outline directions for future research.

1 Introduction

Generative AI systems now produce text, images, audio, and video at scale and at a quality that rivals human-produced content [13]. Authors, journalists, researchers, and institutions increasingly incorporate these tools into their workflows, raising questions about transparency, accountability, and provenance. Who is responsible for AI-assisted content? How much of a document was produced by a machine? Can readers trust what they see?

Attribution badges address these questions by attaching structured signals to content that describe the origin and degree of AI involvement. The concept draws on established traditions in science and digital credentialing. In academic publishing, authorship statements and the Contributor Roles Taxonomy (CRediT) track individual human contributions to papers [16]. In education, digital open badges certify competencies and achievements [25]. In digital media, provenance standards such as C2PA embed cryptographically signed metadata into files [1]. AI attribution badges extend these traditions to the new challenge of hybrid human-AI content.

Despite growing interest, the field lacks a unified taxonomy or standard. Academic publishers have issued disclosure guidelines [18, 29], but these remain largely text-based and self-reported. Technical watermarking approaches offer automated verification [30], but face robustness and privacy challenges [19]. Social media platforms apply simplistic labels that fail to convey the nuance of human-AI collaboration [12]. Regulatory frameworks, most notably the EU AI Act, mandate disclosure for certain AI systems, but implementation details remain in flux.

This survey organizes and synthesizes the existing literature on AI attribution badges. We define the term broadly to include any structured mechanism that signals AI involvement in a content artifact, whether technical (watermarks, provenance manifests), declarative (statements, policies, usage cards), or visual (labels, icons). Our goals are to:

1. Identify the main categories of AI attribution badge approaches and their properties.
2. Compare existing initiatives and standards along key dimensions.
3. Highlight open technical and sociotechnical challenges.
4. Suggest directions for future research and standardization.

The rest of this paper is organized as follows. Section 2 establishes background on attribution in science and digital media. Section 3 presents a taxonomy of AI attribution badge approaches. Section 4 reviews technical mechanisms. Section 5 reviews declarative and policy-based approaches. Section 6 discusses the sociotechnical context. Section 7 identifies open challenges. Section 8 concludes.

2 Background

2.1 Attribution in Scientific Publishing

Scientific publishing has long grappled with questions of credit allocation. The norms of authorship developed over centuries, but multi-author papers have strained the traditional framework [22]. Ghost authorship and gift authorship both misrepresent the true origin of intellectual contributions [15].

The CRediT taxonomy [16] was developed to address this by providing a standardized vocabulary of 14 contributor roles, ranging from conceptualization and methodology to writing and visualization. CRediT allows fine-grained attribution but does not yet accommodate AI systems as contributors. Project Rachel, a study creating and tracking an AI academic identity named Rachel So, demonstrates that AI can publish papers that pass peer review and accumulate citations, exposing gaps in current attribution norms [17].

The prevalence of undisclosed AI-generated text in the literature is significant. One study found that approximately 20% of neurosurgical manuscripts contained sections flagged as AI-generated [21]. The Academ-AI repository documents hundreds of examples of suspected undeclared AI use in journals from major publishers [7]. These findings motivate more robust attribution mechanisms.

2.2 Attribution in Digital Media

Digital content has long faced authenticity challenges. Watermarking, metadata standards, and cryptographic signing have been used to establish provenance for images and documents. The Coalition for Content Provenance and Authenticity (C2PA) standard, developed by a consortium including Adobe, BBC, Microsoft, and others, provides an open technical specification for encoding the origin and editing history of digital media in a cryptographically signed manifest [1, 10].

Social media platforms have adopted varying disclosure strategies. Meta attaches expandable “AI Info” labels, TikTok uses “AI-generated” tags, LinkedIn employs Content Credentials metadata, and YouTube applies broad synthetic content warnings [12]. Research shows that label wording strongly influences user perception: labels that mention AI involvement reduce perceived credibility and sharing willingness, while oversimplified labels suggesting no human involvement may unfairly penalize AI-assisted work [12].

2.3 Digital Badges and Micro-Credentials

Digital open badges, introduced by the Mozilla Foundation in 2012, are electronically verifiable signifiers of achievements or competencies. They consist of a graphical icon linked to metadata that records the recipient, issuing organization, criteria, and date [25]. The Open Badges standard, now maintained by 1EdTech, supports portability and public sharing. Blockchain-based extensions have been proposed to make badge credentials tamper-resistant and decentralized [6].

AI attribution badges can be seen as a specialization of this concept: structured, potentially verifiable signals that communicate AI involvement in content creation. Unlike competency badges, which certify skills of a person, AI attribution badges certify properties of a content artifact.

3 A Taxonomy of AI Attribution Badge Approaches

We organize AI attribution badge approaches along three axes:

Mechanism: How is the badge implemented? Options include technical embedding (watermarks, cryptographic manifests), declarative statements (text in papers, disclosure forms), visual labels (icons, banners), and machine-readable metadata.

Granularity: What level of detail does the badge convey? Options range from binary (AI used / not used) to categorical (AI used for writing, analysis, images) to quantitative (percentage of content generated by AI) to role-based (AI contributed to specific tasks following a taxonomy).

Verifiability: Can the badge be independently verified? Options range from self-declared (author states AI was used) to automatically detected (watermark or classifier) to cryptographically verified (C2PA signature from model provider).

Table 1 summarizes representative approaches along these axes.

Table 1: Representative AI attribution badge approaches.

Approach	Mechanism	Granularity	Verifiability
Journal disclosure statements	Declarative	Binary or categorical	Self-declared
AI Usage Cards [27]	Declarative + metadata	Role-based	Self-declared
CRediT taxonomy [16]	Declarative	Role-based	Self-declared
C2PA content credentials [1]	Cryptographic metadata	Process-based	Cryptographic
Text watermarks [30]	Technical embedding	Binary	Automated
Image watermarks [5]	Technical embedding	Binary	Automated
AI model cards [4]	Documentation	System-level	Self-declared
AI Bill of Materials [26]	Cryptographic metadata	System-level	Cryptographic
Social media labels [12]	Visual	Binary or categorical	Platform-enforced

Each quadrant of this design space presents distinct trade-offs. Technical mechanisms offer automated verifiability but may be brittle, removed, or inapplicable to text. Declarative mechanisms are flexible but depend on author honesty. Cryptographic approaches provide strong assurance but require integration into generation pipelines.

4 Technical Mechanisms

4.1 Watermarking

Watermarking embeds imperceptible signals in content that can later be detected to establish AI origin. The SoK on watermarking for AI-generated content [30] provides a systematic overview of techniques for text, image, audio, and video.

For images, methods such as TrustMark [5] embed hidden patterns that survive compression, cropping, and noise, enabling reliable provenance recovery even after re-sharing. Google DeepMind's SynthID embeds imperceptible watermarks in images, audio, text, and video generated by its models.

For text, token-based watermarking works by partitioning the vocabulary into "green" and "red" token lists, then biasing generation toward the green list [30]. Detection checks whether the proportion of green tokens exceeds a threshold. Adaptive approaches apply watermarking selectively to high-entropy token positions to minimize impact on text quality [30]. However, text watermarks face serious robustness challenges: paraphrase attacks, synonym substitution, and token editing can remove watermarks without substantially degrading text quality [19]. Spoofing attacks can infer hidden watermark signatures and forge watermarked content, potentially implicating innocent authors.

A fundamental limitation of watermarking as an attribution badge mechanism is that it provides only a binary signal (watermarked / not watermarked by a particular model). It does not convey the degree of AI involvement, the nature of human oversight, or which parts of a document were AI-generated versus human-written. It also fails entirely for AI-polished text, where human content is refined by AI without full generation [20].

4.2 Cryptographic Provenance

Cryptographic provenance offers stronger guarantees than statistical watermarks. The C2PA standard [1, 2] uses a manifest structure containing assertions about origin, editing history, and tools used. Assertions are signed by the entity that created or modified the content using public key infrastructure (PKI), similar to HTTPS certificate authorities. The manifest is cryptographically bound to the content through a hash; any modification breaks the binding.

C2PA supports both "hard bindings" (cryptographic hashes of pixel data or raw file bytes) and "soft bindings" (perceptual hashes or watermarks that survive format changes). The standard includes assertions such as `ai_inference`, `ai_training`, and `data_mining`, which can be used by content creators to signal AI involvement and specify training consent.

The AI Bill of Materials (AIBOM) [26] extends the Software Bill of Materials (SBOM) concept to AI models, providing a verifiable record of datasets, model metadata, and training environment details, signed by a neutral third party during training. This addresses attribution at the model level rather than the artifact level.

Cryptographic provenance has significant advantages as an attribution badge mechanism: it is tamper-evident, machine-readable, and expressive. Its main limitations are adoption barriers (generative AI tools must be updated to sign their outputs), fragility to transformations (C2PA signatures break when images are screenshot or heavily edited), and dependence on PKI infrastructure.

4.3 AI Detection

Detection-based attribution identifies AI-generated content after the fact, without relying on signals embedded during generation. Classifiers trained on human-written versus AI-generated text corpora, zero-shot methods based on log probability curvature (DetectGPT), and adversarial detectors (RADAR [9]) all attempt this task.

The MAGE benchmark [13] shows that detection remains challenging across diverse domains and models, particularly in out-of-distribution settings. Detection accuracy is further complicated by AI-polished text, where the degree of AI involvement is gradual rather than binary [20]. As language models improve, the linguistic gap between AI and human text narrows, making reliable detection theoretically and practically harder [19].

Detection cannot form the basis of a reliable attribution badge system for two reasons. First, it produces probabilistic classifications, not verifiable attribution. Second, it can generate false positives, wrongly accusing human authors of AI-generated content, which has been shown to disproportionately affect non-native English speakers.

Despite these limitations, detection tools serve a supporting role in badge ecosystems: they can flag content for further review, motivate proactive disclosure, and help enforce disclosure policies.

5 Declarative and Policy-Based Approaches

5.1 Journal Disclosure Policies

The most widespread AI attribution badges in academic publishing are textual disclosure statements. Following the wide adoption of ChatGPT in late 2022, most major academic publishers issued guidelines requiring authors to declare AI tool use in their submissions. Analysis of publisher policies [18] reveals a broad consensus that: (1) AI tools may not be listed as authors; (2) AI use must be disclosed; (3) authors bear responsibility for accuracy and integrity.

However, policies vary considerably in their requirements. Some journals specify that authors must name the tool used, describe the purpose, and identify the sections affected [29]. Others require only a general statement. Few journals require machine-readable disclosure. A scientific future where AI is embedded in the research process will require policies that go beyond binary disclosure [8].

Undisclosed AI use remains common despite these policies. The Academ-AI database [7] documents cases identifiable by characteristic LLM verbiage (such as "I cannot fulfill this request" appearing in published text) that slipped through editorial review. The prevalence of undetectable undisclosed AI use is likely far higher.

5.2 AI Usage Cards

AI Usage Cards [27] represent a more structured approach to declarative attribution. Proposed as an analog to model cards and datasheets for datasets, they provide a standardized form for reporting AI use across the scientific process. The framework captures three dimensions: transparency (where and how AI was used), integrity (whether humans approved AI-generated content), and accountability (who is responsible).

AI Usage Cards are available in machine-readable formats (LaTeX, XML, JSON, CSV) and can be generated through an online questionnaire at `https://ai-cards.org`. Their phase-based design, aligned with the stages of the scientific process (hypothesis generation, literature review, data collection, analysis, writing), captures more nuance than binary disclosure statements. They can in principle be adapted to domains beyond science.

5.3 Model Cards and Foundation Model Transparency Reports

Attribution badges can target the AI system rather than individual content artifacts. Model cards [4] document a model's intended use, performance characteristics, potential biases, and limitations. The Foundation Model Transparency Index and corresponding transparency reports [4] propose institutionalizing transparency reporting for foundation models, drawing on precedents from social media transparency reports.

While model-level documentation does not directly badge individual pieces of content, it provides essential context for interpreting content-level badges. Knowing which model was used and its documented properties is necessary to assess AI involvement claims.

5.4 The CRediT Taxonomy and AI Contributions

CRediT [16] provides a controlled vocabulary of author contribution roles. Extending CRediT to accommodate AI contributions is a promising direction for AI attribution badges in academic publishing. Such an extension would allow authors to specify not only that AI was used but which tasks AI performed: writing the original draft, creating visualizations, analyzing data, reviewing the manuscript. This approach aligns with the human-AI collaboration disclosure design space explored by [12], who found that role-based timelines and task-based presentations convey more information than simple text labels.

Project Rachel [17] argues for creating a dedicated "AI Author ID" system for AI entities, parallel to the ORCID system for human researchers. Such a system would enable citation practices and impact metrics tailored to AI-generated scholarship.

6 Sociotechnical Context

6.1 User Perception and Trust

Attribution badges are meaningful only if they influence user perception and behavior. Research on AI disclosure in journalism [12] finds that textual disclosures are least effective, while interactive chatbot formats providing in-depth information are most effective. Visual representations matter: role-based timelines amplify perceived AI contribution, while task-based timelines shift perception toward human involvement.

Label wording strongly affects perception. "AI-generated" labels reduce belief in news headlines and sharing willingness, even when humans reviewed the content [12]. Oversimplified labels risk creating a false binary between human and AI content, when in practice most AI-assisted content lies on a spectrum.

Trust in deepfake detection AI affects how users respond to synthetic content warnings [31]. Users calibrate their reliance on AI-based detection tools based on perceived risk and prior AI performance, suggesting that attribution badge systems must be designed with an understanding of how users process and act on disclosure information.

Generative AI content labeled as AI-produced is perceived as comparably credible to human-authored content but less readable [12]. This suggests that attribution badges may not automatically undermine trust in AI-assisted work, provided the disclosure is transparent and the content quality is maintained.

6.2 Regulatory Environment

The EU AI Act mandates transparency for certain AI system categories. Article 50 requires providers of AI systems that interact with humans to disclose that users are interacting with an AI unless the context makes this obvious. Exemptions exist for AI-reviewed content, though user demand for origin information persists even when such exemptions apply [12].

Confidence-building measures proposed in international AI governance contexts [23] include content provenance and watermarking as tools for verifying AI origin and maintaining trust in AI-generated information. These measures are particularly relevant for high-stakes applications in journalism, healthcare, and government.

The regulatory landscape creates heterogeneous compliance requirements across jurisdictions, complicating the design of universal attribution badge standards. Interoperable, jurisdiction-agnostic badge systems would reduce compliance costs while providing consistent information to users.

6.3 Data Provenance and Training Consent

AI attribution badges serve not only readers of AI-generated content but also content creators whose work may have been used to train AI systems. Data provenance standards [14] address the challenge of establishing authentic, consensual, and documented training data. The C2PA standard's `data_mining` assertion allows content creators to signal training consent preferences, integrating attribution badges into the content creation-to-training pipeline.

AI data transparency remains poor across a wide range of deployed systems [28]. Improving attribution for AI-generated content therefore intersects with improving documentation of AI training processes: the AIBoM [26] and Foundation Model Transparency Reports [4] both address this upstream need.

7 Open Challenges

7.1 Verification and Enforcement

The most critical challenge for AI attribution badges is verification. Self-declared badges depend entirely on author honesty. Detection-based verification is unreliable, especially for polished or partially AI-generated content [20, 13]. Cryptographic provenance requires AI tools to sign their outputs at generation time, which is not universally implemented.

Enforcement of disclosure policies by journals is weak. Publishers rely primarily on editorial vigilance and automated detection tools to identify undisclosed AI use, but both methods have high error rates. The academic community has recognized that a shift from detection to transparency is needed [21], but the mechanisms for encouraging and enforcing this shift remain underdeveloped.

7.2 Expressiveness versus Simplicity

Effective attribution badges must balance expressiveness (capturing the degree and nature of AI involvement) with simplicity (being understandable to non-expert users and easy to implement). Binary labels are insufficient for content on the human-AI collaboration spectrum. Role-based or task-based formats convey more information [12] but require more effort from content creators and more cognitive load from readers.

The design of graduated badge systems is an open research problem. What are the right categories? How granular should the disclosure be? How should composite content (partly AI-written, partly human-edited) be represented? These questions require empirical user research as well as normative input from the communities affected.

7.3 Robustness and Adversarial Resistance

Technical attribution mechanisms face adversarial removal. Text watermarks can be eliminated by paraphrasing [19]. Image watermarks can be removed by transformations such as cropping, compression, and screenshots. Even cryptographic C2PA bindings break when images are edited [24]. Soft bindings based on perceptual hashing are more resilient but less precise.

An alternative framing is to watermark authentic human content rather than AI-generated content [11], using hardware-level cryptographic signing at capture time. This approach, analogous to HTTPS certificate authorities, would make human-produced content verifiable without requiring AI systems to brand their outputs. Practical adoption, however, requires integration into cameras and devices at scale.

7.4 Standardization and Interoperability

The attribution badge ecosystem is currently fragmented. Academic publishers, social media platforms, AI model providers, and standards bodies have each developed their own approaches. These are not interoperable: a C2PA credential embedded in an image is invisible to a journal's text-based disclosure system; an AI Usage Card attached to a paper carries no information visible to a social media platform.

A universal, open attribution badge standard analogous to HTTPS or ORCID would reduce fragmentation. The Research Attribution Markup Language (RAML), proposed in [3], is one such initiative targeting AI contributions in academic manuscripts. Whether any single standard can achieve broad adoption across heterogeneous communities and use cases remains an open question.

7.5 Graded Human-AI Collaboration

Perhaps the most conceptually challenging problem is representing graded human-AI collaboration on a continuous spectrum. Current badges treat AI involvement as a category rather than a quantity. But research shows that the human-AI collaboration ratio substantially affects user perception [12], and that role-based or task-based visualizations shift these perceptions in predictable ways.

A badge system for the spectrum would need to distinguish among, for example: AI wrote the first draft and a human made minor edits; a human wrote a draft and AI polished it; a human specified an outline and AI expanded each section; and a human reviewed and corrected a fully AI-generated document. These scenarios have different ethical and epistemic implications, and the attribution badge should reflect those differences.

8 Conclusion

AI attribution badges are a necessary infrastructure for the era of generative AI. They serve multiple functions: informing readers about content origin, protecting human creators from false accusations, enabling accountability for AI systems, and supporting regulatory compliance. This survey has reviewed the main approaches across technical, declarative, and sociotechnical dimensions.

We find that no existing approach is simultaneously expressive, verifiable, interoperable, and simple. Technical mechanisms offer automated verifiability but are brittle and provide only binary signals. Declarative mechanisms are flexible but rely on self-reporting. Cryptographic provenance offers the strongest guarantees but requires deep integration into AI generation pipelines and remains fragile to content transformation.

The most promising near-term path combines role-based declarative badges (extending frameworks like CRediT and AI Usage Cards to AI contributors) with emerging cryptographic provenance standards (C2PA and its successors). In the longer term, hardware-level content signing for authentic human-created content may shift the problem from "labeling AI output" to "verifying human origin," which may be more tractable.

Key open challenges include: standardizing the vocabulary of AI contribution roles; achieving interoperability across platforms and use cases; developing robust verification mechanisms; and designing badges that are informative to diverse user populations without creating undue cognitive load. Progress on these challenges requires collaboration across the research communities of human-computer interaction, machine learning, cryptography, information science, and academic publishing.

Acknowledgments

Generative AI has been used to prepare this paper.

References

[1] Kar Balan, Andrew Gilbert, and John P. Collomosse. Content arcs: Decentralized content rights in the age of generative ai. *ArXiv*, abs/2503.14519, 2025.

[2] Kar Balan, Robert Learney, and Tim Wood. A framework for cryptographic verifiability of end-to-end ai pipelines. *Proceedings of the 2025 ACM International Workshop on Security and Privacy Analytics*, 2025.

[3] Joeran Beel, Min-Yen Kan, and Moritz Baumgart. Evaluating sakana's ai scientist: Bold claims, mixed results, and a promising future? *ACM SIGIR Forum*, 59:1 – 20, 2025.

[4] Rishi Bommasani, Kevin Klyman, Shayne Longpre, Betty Xiong, Sayash Kapoor, Nestor Maslej, Arvind Narayanan, and Percy Liang. Foundation model transparency reports. pages 181–195, 2024.

[5] Tu Bui, S. Agarwal, and J. Collomosse. Trustmark: Universal watermarking for arbitrary resolution images. *ArXiv*, abs/2311.18297, 2023.

[6] Alberto Francia, Stefano Mariani, Giuseppe Adduce, Sandro Vecchiarelli, and Franco Zambonelli. Digital management of competencies in web 3.0: The c-box® approach. *Future Internet*, 15:350, 2023.

[7] Alex Glynn. Academ-ai: documenting the undisclosed use of generative artificial intelligence in academic publishing. 2024.

[8] Renee Hoch and J. Clarke. A scientific future shared with ai. *PLOS Biology*, 23, 2025.

[9] Xiaomeng Hu, Pin-Yu Chen, and Tsung-Yi Ho. Radar: Robust ai-text detection via adversarial learning. *ArXiv*, abs/2307.03838, 2023.

[10] Shrey Jain, Zoe Hitzig, and Pamela Mishkin. Position: Contextual confidence and generative ai. *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 281–301, 2023.

[11] Houssam Kherraz. On-device watermarking: A socio-technical imperative for authenticity in the age of generative ai. *ArXiv*, abs/2504.13205, 2025.

[12] Amber Kusters, Pooja Prajod, Pablo Cesar, and Abdallah El Ali. More human or more ai? visualizing human-ai collaboration disclosures in journalistic news production. *ArXiv*, abs/2601.11072, 2026.

[13] Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. Mage: Machine-generated text detection in the wild. pages 36–53, 2023.

[14] Shayne Longpre, Robert Mahari, Naana Obeng-Marnu, William Brannon, Tobin South, Katy Ilonka Gero, Sandy Pentland, and Jad Kabbara. Data authenticity, consent, & provenance for ai are all broken: what will it take to fix them? pages 32711–32725, 2024.

[15] Abdelghani Maddi and J. A. Teixeira da Silva. Beyond authorship: Analyzing contributions in plos one and the challenges of appropriate attribution. *Journal of Data and Information Science*, 9:88 – 115, 2024.

[16] N. Mekeko and E. Tikhonova. Description of author contributions while submitting a manuscript: The credit system. *Journal of Employment and Career*, 2022.

[17] Martin Monperrus, Benoit Baudry, and Clément Vidal. Project rachel: Can an ai become a scholarly author? *ArXiv*, abs/2511.14819, 2025.

[18] Mike Perkins and Jasper Roe. Academic publisher guidelines on ai usage: A chatgpt supported thematic analysis. *F1000Research*, 12, 2023.

[19] Vinu Sankar Sadasivan, Aounon Kumar, S. Balasubramanian, Wenxiao Wang, and S. Feizi. Can ai-generated text be reliably detected? *ArXiv*, abs/2303.11156, 2023.

[20] Shoumik Saha and S. Feizi. Almost ai, almost human: The challenge of detecting ai-polished writing. *ArXiv*, abs/2502.15666, 2025.

[21] Daniel Schneider, Akash Mishra, Jacob Gluski, Harshal A Shah, Max Ward, Ethan D. L. Brown, Daniel M. Sciubba, and S. Lo. Prevalence of artificial intelligence-generated text in neurosurgical publications: Implications for academic integrity and ethical authorship. *Cureus*, 17, 2025.

[22] Huawei Shen and A. Barabási. Collective credit allocation in science. *Proceedings of the National Academy of Sciences*, 111:12325 – 12330, 2014.

[23] Sarah Shoker, A. Reddie, Sarah Barrington, Miles Brundage, H. Chahal, Michael Depp, Bill Drexel, Ritwik Gupta, Marina Favaro, J. Hecla, Alan Hickey, Margarita Konaev, K. Kumar, Nathan Lambert, A. Lohn, Cullen O'Keefe, Nazneen Rajani, M. Sellitto, Robert Trager, L. Walker, Alexa Wehsener, and Jessica Young. Confidence-building measures for artificial intelligence: Workshop proceedings. *ArXiv*, abs/2308.00862, 2023.

[24] Shree Singhi, Aayan Yadav, Aayush Gupta, Shariar Ebrahimi, and Parisa Hassanizadeh. Provenance detection for ai-generated images: Combining perceptual hashing, homomorphic encryption, and ai detection models. *ArXiv*, abs/2503.11195, 2025.

[25] Natasja Steenkamp, Richard Fisher, and Trevor Nesbit. Understanding accounting students' intentions to use digital badges to showcase employability skills. *Accounting Education*, 33:906 – 934, 2023.

[26] Wiebe Vandendriessche, Jordi Thijsman, Laurens D'hooge, B. Volckaert, and Merlijn Sebrechts. Aibomgen: Generating an ai bill of materials for secure, transparent, and compliant model training. *ArXiv*, abs/2601.05703, 2026.

[27] Jan Philip Wahle, Terry Ruas, Saif M. Mohammad, Norman Meuschke, and Bela Gipp. Ai usage cards: Responsibly reporting ai-generated content. *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 282–284, 2023.

[28] Sophia Worth, Ben Snaith, Arunav Das, Gefion Thuermer, and E. Simperl. Ai data transparency: an exploration through the lens of ai incidents. *ArXiv*, abs/2409.03307, 2024.

[29] Jina Yoo. Defining the boundaries of ai use in scientific writing: A comparative review of editorial policies. *Journal of Korean Medical Science*, 40, 2025.

[30] Xuandong Zhao, Sam Gunn, Miranda Christ, Jaiden Fairoze, Andres Fabrega, Nicholas Carlini, Sanjam Garg, Sanghyun Hong, Milad Nasr, Florian Tramèr, Somesh Jha, Lei Li, Yu-Xiang Wang, and D. Song. Sok: Watermarking for ai-generated content. *2025 IEEE Symposium on Security and Privacy (SP)*, pages 2621–2639, 2024.

[31] Yingfan Zhou, Ester Chen, Manasa Pisipati, Aiping Xiong, and Sarah M. Rajtmajer. Effect of ai performance, risk perception, and trust on human dependence in deepfake detection ai system. *Proceedings of the ACM on Human-Computer Interaction*, 9:1 – 24, 2025.