AI Watermarks: Public Reactions, Technical Evidence, Policy Debate, and Implications for Universities

Executive summary

AI watermarking has moved from a technical research topic into an institutional governance issue. The immediate trigger for the current debate is the application of Article 50 of the EU AI Act from August 2, 2026. It requires providers of generative AI systems to make synthetic audio, image, video, and text outputs detectable in machine-readable form, subject to technical feasibility. A limited transition until December 2, 2026 applies to the marking and detection obligations for systems placed on the market before August 2. The European Commission's Code of Practice offers one route to demonstrating compliance. [1]

The LinkedIn material supplied for this analysis captures this debate during August 2026 across five discussion contexts: Keith King's post on visible versus invisible provenance, Melia Robinson's poll about Claude, Marek Fisera's technical discussion, Guillaume Meyer's critique, and Michael Wade's explanatory FAQ. The sample contains 20 substantive top-level comments that provide enough material for structured qualitative coding, although not for population-level inference about LinkedIn users. 毅filecite孳turn0file0傢

My binary coding, required by the research brief, classifies 12 of 20 comments, 60%, as predominantly resistant or critical of watermarking, and 8 of 20, 40%, as predominantly supportive. The distinction needs care. Much of the "negative" cluster supports transparency while opposing binary, invisible, context-free watermarking. Much of the "positive" cluster supports disclosure, verification, or professional accountability rather than endorsing every watermarking technique. The dominant cleavage therefore is not transparency versus secrecy. It is whether a watermark supplies enough information to support the judgments institutions are likely to make from it. 毅filecite孳turn0file0傢

This interpretation gains support from Melia Robinson's 66-vote poll. Asked whether respondents would continue writing with Claude after watermarking, 42% answered "Yes," 25% "No," and 31% "It depends." The poll therefore looks less resistant than the comment corpus. The difference suggests comment self-selection: people with nuanced concerns have more reason to write a comment than people who simply vote yes. The poll is small and non-random, so this is an interpretive signal rather than an estimate of wider public opinion. 毅filecite孳turn0file0傢

The strongest recurring concern is the distinction between provenance and authorship. A machine-readable mark can indicate that an AI system processed a piece of content. It cannot, by itself, establish who supplied the ideas, who performed the research, how

much the human revised the output, whether sources were checked, whether the use complied with policy, who owns intellectual responsibility, or whether plagiarism occurred. This distinction appears repeatedly in comments by Shawn Kohrman, JURINEX AI, Brad Larsen, Joost de Kinderen, and Cherif Diouf. 毅filecite孳turn0file0傢

The technical literature supports both sides of the argument. Statistical text watermarking is feasible. The foundational Kirchenbauer et al. method embeds a detectable statistical signal by influencing token selection, while Google DeepMind's SynthID-Text demonstrated production-scale watermarking and reported no statistically significant change in user feedback across approximately 20 million Gemini responses. At the same time, detectability varies with text length, entropy, editing, paraphrasing, and implementation. A watermark therefore provides evidence of model processing under specified conditions, not a complete account of authorship or research integrity. [2]

The wider provenance ecosystem is already converging on layered mechanisms. C2PA provides cryptographically signed provenance manifests associated with digital assets. Google has deployed SynthID across several media types. OpenAI currently uses C2PA metadata plus SynthID for supported images and SynthID for supported audio. OpenAI explicitly presents these as complementary provenance signals rather than a single definitive solution. [3]

For universities, the main policy implication is clear: do not convert "watermark detected" into "academic misconduct detected." That inference is invalid. Academic policy should instead separate at least five questions: Was AI involved? What did AI do? Was the use permitted? Did the human verify and take responsibility for the work? Were sources and contributions disclosed appropriately? Current publishing guidance already follows this accountability model. ICMJE, for example, requires disclosure of AI-assisted technologies, excludes AI systems from authorship because they cannot assume responsibility, and leaves responsibility for accuracy, originality, citation, and plagiarism with human authors. [4]

For a university of economics, watermarking offers meaningful value in provenance-sensitive settings such as official communications, synthetic teaching media, image and audio assets, research artifacts, and evidentiary material. Its value is weaker when institutions try to use it as a proxy for student authorship, plagiarism, research quality, or misconduct. A sound policy therefore combines machine-readable provenance with explicit human disclosure, version history, assessment design, source verification, and due process.

Scope, data, and coding method

Corpus and collection limits

The supplied LinkedIn material served as the primary social-data corpus. It contains recent public posts, articles, a poll, and comments concerning AI watermarks and related provenance mechanisms. The material includes identifiable professional roles,

relative posting dates, and reaction counts where LinkedIn exposed them in the captured material. 毅filecite孳turn0file0傢

The source set covers five main discussion contexts:

Discussion source	Self-described role	Date in supplied material	Main topic	Available engagement
Keith King	Former senior US government communications engineer, strategic advisor	Recent, relative date not captured for post	Visible watermark removal versus invisible provenance	Post metrics unavailable
Melia Robinson	Senior Correspondent, Legal and Tech, Business Insider	"5 days ago"	Whether professionals will continue writing with Claude	Poll: 66 votes
Marek Fisera	Marketer, coder, AI practitioner	August 13, 2026	How statistical text watermarking works and how robust it is	Article metrics unavailable
Guillaume Meyer	Founder, Memo	August 16, 2026	Case against invisible statistical text watermarking	Author reports large external engagement, independently unverified here
Michael Wade	Professor at IMD, digital and AI transformation	August 12, 2026	Technical and policy FAQ on Claude watermarking	Comment reactions available for several comments

Source: supplied LinkedIn capture. 斀filecite孳turn0file0傢

There is an important name correction. The requested comparison refers to "Mark Wade," but the supplied LinkedIn article identifies its author as Michael Wade. IMD independently identifies Michael Wade as Professor of Strategy and Digital and Director of its Global Center for Digital and AI Transformation. I therefore compare Keith King with Michael Wade. I found no evidence connecting a Mark Wade to the supplied IMD watermarking article. 斀filecite孳turn0file0傢 [5]

LinkedIn does not provide a stable, complete public research interface for these discussions in this collection environment. I therefore treat the supplied capture as a purposive snapshot rather than a complete scrape. Missing comments, deleted comments, sorting effects, personalized ranking, hidden reaction counts, edited profiles, and LinkedIn's interface can all affect what appears in such a snapshot. The analysis cannot estimate the prevalence of attitudes among LinkedIn users, AI users, academics, lawyers, or the general public.

Coding rules

I treated 20 substantive top-level comments as the unit of analysis. I excluded post bodies from the comment frequency statistics and excluded brief replies by the original post author, such as Michael Wade's "Agreed!", from the N = 20 reaction corpus. I used those author replies only when interpreting the author's own position. 斀filecite孳turn0file0傢

The requested positive-negative clustering creates a methodological problem because several respondents favor transparency while resisting watermarking as the mechanism. I therefore used a forced two-family classification with a secondary strength code:

Positive/supportive includes comments whose dominant evaluation accepts watermarking, disclosure, or the transparency direction without an explicit rejection of the mechanism. +1 indicates cautious or conditional support; +2 indicates clear support.

Negative/resistant includes comments whose dominant evaluation rejects the mechanism, questions its usefulness, stresses harmful inference, or proposes circumvention or replacement. -1 indicates conditional or analytical resistance; -2 indicates strong rejection.

Theme coding was multi-label. One comment can therefore contribute to several themes. Theme counts should not sum to 20.

Coded LinkedIn comment corpus

Commenter	Role from profile	Date shown	Reactions shown	Dominant coded reaction	Strength	Condensed content
Shawn Kohrman	Author, AI judgment and decision design	2d	NA	Negative	-1	Provenance and authorship differ; accountability belongs to a named person
Jasmine Singh	General Counsel, legal AI	5d	2	Positive	+1	Watermarks may expose poor verification, but create privilege and chilling-use questions
Nathan Walter, comment A	Legal-tech CEO	5d	1	Negative	-1	Significance depends on court disclosure rules; method does not determine legal

Commenter	Role from profile	Date shown	Reactions shown	Dominant coded reaction	Strength	Condensed content correctness
Nathan Walter, comment B	Legal-tech CEO	5d	0	Negative	-1	Suspected AI use may lead to heightened scrutiny of citations
Jessica Nguyen	Chief Strategy and Legal Officer	5d	2	Positive	+1	Comfortable with watermarking if professional verification is built into legal workflows
JURINEX AI	Legal AI organization	22h	0	Negative	-1	"Was AI used?" differs from "Was the work professionally verified?"
Michael Drapkin	Patent attorney, AI and patents	4d	1	Positive	+2	Comfortable with visible AI

Commenter	Role from profile	Date shown	Reactions shown	Dominant coded reaction	Strength	Condensed content
						involvement because responsibility and client value matter more
Haydn Jones	Legal-tech strategy executive	5d	0	Negative	-2	Acceptability depends on use case; watermark can stigmatize legitimate use and can be bypassed
Itamar Rosen	Chief Legal Officer, AI/data governance	4d	0	Negative	-1	Questions whether pervasive AI labeling would become excessive
Marek Fisera	Marketing, coding, AI	4d	NA	Positive	+1	Would continue using AI for researc

Commenter	Role from profile	Date shown	Reactions shown	Dominant coded reaction	Strength	Condensed content h, ideas, and fact-validation stages
Brad Larsen	Enterprise growth and operations leader	1d	NA	Negative	-2	AI assistance does not erase human direction, authorship, or accountability
Kenneth B.	CFO	5d	6	Positive	+1	Sees watermarking as consequential for education, exams, and professional documents
Joost de Kinderen	Healthcare executive	4d	2	Negative	-1	Supports transparency but calls the watermark too binary;

Commenter	Role from profile	Date shown	Reactions shown	Dominant coded reaction	Strength	Condensed content
						proposes materiality-based disclosure
Sharon Chisholm	Emergency-management specialist	5d	3	Negative	-2	Expresses strong concern about token-level imprinting from ordinary language assistance
Cherif Diouf	IT, software architecture, AI	4d	2	Negative	-1	Processing signal cannot establish human authorship or responsibility without workflow context
Hartmut Hübner, PhD	AI consultant	2d	NA	Positive	+1	Tool disclosure signals

Commenter	Role from profile	Date shown	Reactions shown	Dominant coded reaction	Strength	Condensed content
						integrity more than hiding assistance does
Łukasz Lech	Professional role not specific enough to classify	2d	1	Negative	-1	Questions whether a tiny correction can mark a much larger human-authored text
Gergely N.	Culture and employee engagement consultant	4d	1	Positive	+2	Explicitly supports clear marking of AI-generated content
Gareth T.	Portfolio manager and financial advisor	5d	2	Negative	-2	Treats cross-model rewriting as an easy way to evade the signal

Commenter	Role from profile	Date shown	Reactions shown	Dominant coded reaction	Strength	Condensed content
David McLoughlin	Chief architect	4d	2	Positive	+2	Calls watermarking a move in the right direction

NA means the supplied capture did not provide a usable reaction count. An absent number was not converted to zero. Source: supplied LinkedIn material.

Among comments with visible metrics, the captured reactions sum to at least 25. Positive-coded comments account for at least 14 of these visible reactions and negative-coded comments for at least 11. These totals have little inferential value because reaction availability is incomplete, LinkedIn reaction counts change over time, comments have different exposure, and the highest reaction count is only six.

The role distribution is also important. Legal professionals and legal-tech executives are unusually prominent in the Robinson discussion. Industry executives and technology professionals dominate the rest. No commenter in the coded N = 20 sample represents a clean sample of university faculty opinion, although Michael Wade is an academic post author. This composition helps explain the strong focus on accountability, evidentiary standards, professional verification, client trust, and legal consequences.

LinkedIn reaction clusters

Cluster comparison

The forced binary clustering produces a 60:40 critical-to-supportive split.

Dimension	Positive/supportive cluster	Negative/resistant cluster
Frequency	8 of 20, 40%	12 of 20, 60%
Strength distribution	3 strong, 5 cautious	4 strong, 8 moderate or conditional
Typical stance	Transparency is useful; AI use is acceptable if	A watermark is too binary, incomplete, stigmatizing, or easy to misinterpret

Dimension	Positive/supportive cluster professionals verify and own the output	Negative/resistant cluster
Central question	"How can provenance improve trust and accountability?"	"What does a watermark prove, and what will institutions wrongly infer from it?"
Human responsibility	Watermark can support stronger checking and disclosure	Human accountability matters more than whether a model touched the text
View of legitimate AI assistance	Compatible with professional work	Often compatible with professional work, but watermarking mischaracterizes the contribution
View of institutional use	Useful signal in appropriate workflows	Dangerous if converted into proof of authorship, misconduct, or low quality
Main underlying values	transparency, traceability, professional integrity, evidentiary confidence	authorship, autonomy, proportionality, procedural fairness, contextual judgment
Representative comments	Gergely N.: clear AI marking is good. Drapkin: AI use is acceptable when experts remain responsible. Jessica Nguyen: verify outputs and incorporate this into professional guidelines.	Shawn Kohrman: provenance differs from authorship. Joost de Kinderen: disclose material AI contribution rather than reduce use to a binary signal. Gareth T.: rewriting can undermine detection.

Source and classifications: analysis of supplied LinkedIn material.
毅filecite孳turn0file0傢

A crucial result is the amount of agreement across the two clusters. Both sides repeatedly reject the proposition that AI involvement removes human responsibility. Positive respondents tend to say, "mark the use and still hold the professional

accountable." Negative respondents tend to say, "hold the professional accountable rather than treating the mark as an authorship judgment." 毅filecite孳turn0file0傢

This makes the debate less polarized than a positive-negative label suggests.

Theme frequencies

The most frequent theme is human verification and accountability, appearing in 9 of 20 comments. Context and materiality of AI use and transparency/trust each appear in 8. Stigma, chilling effects, or false inference appears in 7. Five discuss legal or assessment consequences, three defend AI as legitimate assistance or a productivity tool, and two explicitly discuss easy circumvention or fragility. Because a comment can receive multiple codes, totals exceed N = 20. 毅filecite孳turn0file0傢

Bar chart of theme frequencies in the LinkedIn comment sample

Chart file: PNG

The chart shows why "pro-watermark versus anti-watermark" is an incomplete interpretation. The central issue is governance. Respondents repeatedly ask what decision follows from detection and who remains responsible after AI assistance. 毅filecite孳turn0file0傢

The poll and comments tell different stories

Melia Robinson's poll provides an informative contrast. Among 66 votes, 42% said they would still write with Claude, 25% said they would not, and 31% selected "It depends." Rounding accounts for the reported percentages summing to 98 rather than 100. 毅filecite孳turn0file0傢

The poll therefore has three noteworthy properties.

First, outright refusal is the smallest response category. Second, almost one-third selects conditional acceptance. Third, the comments are more critical than the poll distribution. The likely interpretation is selection rather than contradiction. A voter who sees no issue can click "Yes" and move on. A professional worried about privilege, authorship, verification, stigma, or court rules has more material to explain in a written comment.

This matters for social-media research. Counting comments alone would overstate resistance relative to the accompanying poll. Counting poll votes alone would conceal the institutional reasons for resistance. Combining quantitative poll data with qualitative comments gives the stronger reading.

What drives support and resistance

Why people support AI watermarks

Supporters emphasize transparency. They want audiences, organizations, courts, platforms, or publishers to have some evidence about how digital content originated. That concern has become more salient as synthetic media becomes easier to produce and harder to distinguish visually or stylistically. EU policymakers explicitly frame Article 50 transparency rules around risks of deception and manipulation and the integrity of the information environment. [6]

A second argument is verification. Jasmine Singh's legal example captures the logic: if a watermark encourages lawyers to check citations and quoted authorities before filing, it can reinforce a responsibility the lawyer already has. Jessica Nguyen similarly links acceptance of AI use with explicit verification obligations. The watermark serves as a prompt for scrutiny rather than a judgment about whether the work is good. 毅filecite孳turn0file0傢

Third, some supporters reject the idea that disclosure is stigmatizing. Michael Drapkin's position is essentially that professional value comes from expert judgment and results, so disclosure of AI assistance need not undermine credibility. Hartmut Hübner similarly treats openness about tools as evidence of integrity. 毅filecite孳turn0file0傢

Fourth, supporters see machine-readable provenance as part of digital trust infrastructure. This is strongest for images, audio, video, official records, and other artifacts in which proving origin can matter independently of authorship. C2PA, for example, defines signed manifests that bind provenance assertions and signatures into a verifiable Content Credential. It addresses the history and origin of an asset rather than making a truth judgment about the content itself. [7]

Finally, production evidence suggests text watermarks do not inherently require major quality degradation. In the 2024 SynthID-Text study, Google DeepMind researchers tested watermarked and unwatermarked responses in roughly 20 million live Gemini interactions and found the differences in thumbs-up and thumbs-down rates statistically insignificant. This supports the feasibility case, although results for one watermarking system do not automatically establish the performance of every implementation. [8]

The underlying supportive values are therefore traceability, transparency, professional responsibility, deterrence of deceptive representation, and confidence in digital provenance.

Why people resist AI watermarks

The strongest objection concerns category error: AI processing is not the same as AI authorship.

A researcher can write an article and ask an AI system to correct grammar. A student can develop the argument, gather the evidence, write the initial draft, and use AI for

editing. Another student can ask a model to produce an entire essay from a one-sentence prompt. A binary signal that says a model processed both documents does not distinguish those workflows. This is the point Joost de Kinderen captures with his call for disclosure based on material contribution. 毅filecite孳turn0file0傢

The second objection concerns accountability. Shawn Kohrman's formulation is especially useful: "A watermark tells you what made a file. Only a named person tells you who answers for it." 毅filecite孳turn0file0傢 The academic analogue is strong. A professor's article remains the professor's responsibility even after AI language editing. A student's submission remains the student's responsibility under the assessment rules. The mark does not answer whether the human checked citations, understood the analysis, selected appropriate methods, or accepts responsibility for errors.

The third objection concerns false inference. A technically accurate statement such as "this text contains a statistical pattern associated with model X" can become an institutionally inaccurate statement such as "the student did not write this." Those propositions are different. Watermarking creates the risk of an inference gap between what a detector technically measures and what a decision-maker believes it means.

The current debate around generic AI-writing detection already illustrates why evidence thresholds matter. Turnitin states explicitly that its AI-writing model can misidentify human, AI-generated, and AI-paraphrased text and says the result should not serve as the sole basis for adverse action against a student. Watermark detection is technically different from stylistic AI detection, so their error properties should not be conflated. The governance principle still transfers: a probabilistic or partial technical signal needs human evaluation and policy context. [9]

Fourth, respondents fear stigma and chilling effects. Haydn Jones distinguishes pleading, email, and LinkedIn use because audiences attach different social meanings to AI involvement. Brad Larsen objects to a signal that can cause observers to discount work even when a human directed, verified, and owns the result. Joost de Kinderen similarly fears that a binary mark encourages readers to infer more than the evidence warrants. 毅filecite孳turn0file0傢

Fifth, text watermarks face robustness limits. Statistical watermarking works by influencing token selection. The original Kirchenbauer et al. framework uses randomized preferred token sets, while SynthID-Text uses a tournament-based sampling mechanism. Detection accumulates statistical evidence across tokens, so length, entropy, transformations, paraphrasing, and model implementation matter. [2]

That creates an asymmetry noted in the LinkedIn debate. An uninformed user may preserve a watermark; a motivated actor can alter text substantially, use another model, or rewrite it manually. Haydn Jones and Gareth T. explicitly raise this point. 毅filecite孳turn0file0傢 A provenance mechanism can still have value despite imperfect robustness, but universities should not design high-stakes disciplinary systems on the assumption that all AI use is equally detectable.

Sixth, users object to invisible vendor-controlled signaling. Transparency for the reader and detectability for an institution are not the same property. An invisible mark that only a provider or authorized detector can read introduces questions about interoperability, auditability, access to keys, evidence preservation, appeal rights, and vendor dependency.

The underlying resistant values are therefore contextual authorship, personal autonomy, proportionality, fairness, due process, legitimate use of productivity tools, and protection against overinterpretation.

The debate is better represented as a sequence of questions

A defensible governance model separates five questions:

1. Did an AI system process this content?
2. What kind of processing occurred?
3. How material was the AI contribution?
4. Did the human verify, understand, and take responsibility for the result?
5. Was that use permitted and properly disclosed under the relevant rules?

Watermarking can contribute strongly to the first question in some implementations. It contributes weakly to questions two and three. It does not answer questions four and five.

That distinction explains most of the disagreement in the LinkedIn corpus.
毅filecite孳turn0file0傣

Keith King and Michael Wade

Keith King's position

Keith King frames watermarking primarily as a provenance and infrastructure problem. His post contrasts visible disclosure with invisible technical provenance and argues that the industry is moving toward "layered provenance rather than relying on a single visible marker." 毅filecite孳turn0file0傣

His emphasis falls on three levels.

At the human level, a visible mark allows a viewer to notice AI origin without specialist tools.

At the machine level, invisible mechanisms can survive when visible marks are absent and support automated verification.

At the governance level, platforms, governments, enterprises, media organizations, and technical standards need infrastructure that allows content origin to be authenticated.
毅filecite孳turn0file0傣

This position closely resembles the layered provenance direction visible in current industry practice. OpenAI, for example, combines C2PA metadata with SynthID for supported images and uses SynthID for supported audio, describing provenance as a multi-layered problem. [10] C2PA itself treats signed provenance claims and their verification as the core mechanism. [7]

King is therefore best characterized as provenance-oriented and infrastructure-oriented rather than as an advocate of a single universal visible watermark.

Michael Wade's position

Michael Wade approaches the topic differently. His August 12 FAQ focuses on how text watermarking works, which systems are affected, what detection means, how editing affects the signal, how different providers compare, and what institutional users should infer from a future detection result. 毅filecite孳turn0file0傣

Wade is receptive to the transparency objective. His reply "Agreed!" to a commenter who wrote that clear marking of AI-generated content is a good thing provides direct evidence of that stance. Yet his FAQ also places clear limits on evidentiary interpretation. The most important statement for universities is his view that a watermark hit should serve as one input to human judgment rather than a standalone verdict of cheating, plagiarism, or deception. 毅filecite孳turn0file0傣

Wade therefore occupies a middle position. He accepts watermarking as an emerging transparency and compliance mechanism but recognizes that model processing cannot establish authorship.

Where they differ

Dimension	Keith King	Michael Wade
Primary frame	Digital provenance and trust infrastructure	Technical explanation and operational consequences
Central distinction	Human-visible disclosure versus machine-readable provenance	AI processing versus authorship and interpretation
Preferred architecture	Layered provenance	Watermarking as one imperfect but usable signal
Main institutional audience	Governments, platforms, media, enterprises, security-oriented actors	Professionals, organizations, educators, AI users

Dimension	Keith King	Michael Wade
Central risk	Loss of trust when synthetic and authentic media become difficult to distinguish	Misreading what a statistical watermark does and does not establish
Position on transparency	Strongly favorable to provenance mechanisms	Favorable, with explicit interpretive limits
Implication for universities	Build reliable provenance infrastructure	Never turn detection into an automatic misconduct judgment

Sources: supplied posts and comments. 豰filecite孳turn0file0傢

Their positions are more complementary than contradictory. King addresses authenticity of artifacts. Wade addresses interpretation of AI involvement. A university needs both layers.

King's frame is strongest when the question is, "Where did this digital object come from, and has its provenance survived?"

Wade's frame is strongest when the question is, "What does AI involvement mean for authorship, responsibility, and policy?"

This distinction maps directly onto research governance. A cryptographically signed provenance chain can help establish the origin of a dataset, figure, audio file, video, or institutional document. A contribution statement must still establish who conceived the study, chose variables, ran models, interpreted results, checked evidence, and approved publication.

Evidence beyond LinkedIn

Regulation now gives watermarking institutional significance

Article 50 of the EU AI Act applies from August 2, 2026. The Commission states that providers must meet transparency obligations concerning AI-generated content and that machine-readable marking and detection form part of those obligations. Systems placed on the market before August 2 receive a limited transition for Article 50(2) marking and detection until December 2, 2026. [11]

The Commission's Code of Practice distinguishes provider-side marking from deployer-side labeling. Providers of systems that generate synthetic text, audio, images, or video face machine-readable marking requirements. Professional deployers face separate disclosure duties for deepfakes and certain AI-generated or manipulated text published to inform the public on matters of public interest. The Commission also

recognizes human review and editorial responsibility in the public-interest text regime. [12]

This last point is important for the authorship debate. EU governance itself does not reduce every AI-assisted publication to the same category. Human review, editorial control, use context, provider duties, and deployer duties matter. [13]

Potential penalties give institutions a reason to follow implementation closely. Commission guidance states that relevant fines can reach €15 million or 3% of worldwide annual turnover for the preceding financial year, with proportionality considerations for smaller firms. [14]

Text watermarking has moved beyond proof of concept

The 2023 Kirchenbauer et al. ICML paper established an influential statistical approach to text watermarking. The method partitions candidate tokens into randomized categories and slightly favors selected tokens during generation. Accumulated token choices then yield a statistical signal that a detector can test. The authors reported efficient detection with negligible quality impact in their experiments. [15]

Google DeepMind's SynthID-Text advanced this approach to production scale. Its Nature paper describes tournament sampling, efficient detection, compatibility with production inference techniques, and large-scale deployment in Gemini. Its approximately 20 million-response experiment found no statistically significant difference in user feedback between watermarked and comparison responses. [8]

The paper also makes an important conceptual point for policy. Detection is statistical. Performance depends on the amount and properties of text available for analysis. A university therefore needs calibrated thresholds, documented detector versions, error estimates, and evidentiary procedures if it ever incorporates such signals into formal processes. [8]

Provenance is broader than text watermarking

C2PA approaches the problem from another direction. A Content Credential is a cryptographically signed provenance manifest containing assertions about an asset. Its trust model centers on signatures, credentials, content bindings, and validation rather than linguistic style. [7]

This gives C2PA several strengths for institutional artifacts. A university can associate provenance information with digital files, preserve provenance across workflows that support the standard, and verify whether signed claims remain valid. Yet provenance does not establish factual truth. A correctly signed synthetic image can still depict something false; a genuine photograph can appear in a misleading context. Provenance supports verification of origin, not automatic verification of meaning.

The distinction mirrors the LinkedIn comments: origin, authorship, truth, quality, and accountability are separate properties.

Current industry direction favors multiple signals

OpenAI's current provenance implementation illustrates a layered model. Supported images contain C2PA metadata and SynthID watermarks. Supported OpenAI-generated audio includes SynthID, and OpenAI's public verification system checks supported provenance signals. [16]

As of August 18, 2026, OpenAI's official provenance documentation lists supported provenance signals for images and audio rather than text. This makes a useful comparison with Gemini's productionized SynthID-Text and the current discussion around Claude text watermarking. [17]

Recent reporting on Claude shows why the issue has become prominent this month. Coverage reports that Anthropic is moving toward invisible machine-readable marking for Claude text and C2PA-style provenance for images, with the text mechanism linked to SynthID-Text and EU transparency requirements. The reporting also documents disagreement about whether token-level influence changes prose quality and whether users will interpret AI-processing signals too broadly. [18]

Scholarship is shifting toward process transparency

A July 2026 paper by Germani and Spitale directly addresses the problem seen in the LinkedIn corpus. The authors argue that converting AI-generation signals into binary labels can conflate machine involvement with truthfulness and can misrepresent human-AI co-creation. They advocate process transparency and information literacy as complementary or preferable responses in settings where authorship and meaning matter. The paper is recent and, at present, should be treated as part of an active scholarly debate rather than settled consensus. [19]

Academic-publishing guidance already focuses on accountable human processes. ICMJE requires authors to disclose whether and how AI-assisted technologies contributed to submitted work. It excludes AI systems from authorship because authors must assume responsibility for accuracy, integrity, and originality. Humans remain responsible for reviewing generated material, preventing plagiarism, checking attribution, and verifying citations. [4]

Higher education is simultaneously moving toward formal AI governance. UNESCO reported in September 2025 that nearly two-thirds of surveyed institutions associated with UNESCO Chairs or UNITWIN networks either had AI guidance or were developing it. UNESCO's wider guidance emphasizes human agency, privacy, equity, pedagogical suitability, policy development, and institutional capacity rather than a detection-only response. [20]

The external evidence therefore supports a layered interpretation:

Question	Best-suited evidence
Did a specific AI system generate or process content?	Watermark signal, where available
What is the file's declared technical provenance?	C2PA or comparable signed metadata
Did the student or researcher plagiarize?	Source comparison, citation analysis, investigation
Who authored the intellectual contribution?	Draft history, contribution records, oral explanation, research records
Did AI use comply with course or journal policy?	Policy plus disclosure and workflow evidence
Is the content factually correct?	Independent verification and scholarly review
Who is accountable?	Named human author, student, researcher, editor, or responsible office

No single technical signal answers all seven questions.

Implications and recommendations for a university of economics

For a university of economics, the watermarking debate matters more than a narrow plagiarism-detection discussion suggests. Economics, finance, marketing, management, accounting, statistics, and business analytics all involve mixtures of prose, quantitative analysis, data, code, presentation materials, visualizations, reports, and increasingly AI-supported workflows. Watermark policy will therefore affect research, teaching, student assessment, publishing, data governance, and administration.

Research integrity

A watermark can help document that a model processed a research artifact, but it cannot establish whether the researcher designed the study, selected the model correctly, fabricated data, misunderstood causal assumptions, misreported a coefficient, checked citations, or interpreted the analysis responsibly.

The university should therefore define research integrity around attributable human responsibility rather than around the presence or absence of an AI signal. This aligns with current publishing guidance, which keeps human authors responsible for accuracy, originality, citation, and disclosure even when they use AI tools. [4]

For empirical economics and business research, a stronger provenance record would include the raw-data source, transformations, code version, analytical environment, model and package versions, AI tools used, material AI contributions, human

verification steps, and final human sign-off. A text watermark can become one additional provenance field, but should never replace this record.

A practical disclosure taxonomy would classify AI use by function:

AI contribution	Example	Recommended treatment
Language correction	Grammar and style correction of human draft	Brief disclosure if required by journal or university
Translation	Translating human-authored material	Disclose tool and human verification where material
Research assistance	Search terms, literature leads, brainstorming	Record when relevant; verify all sources independently
Analytical assistance	R/Python code generation, debugging, formula explanation	Preserve code history and independently validate output
Draft generation	AI produces paragraphs or sections	Material disclosure
Quantitative interpretation	AI interprets results or proposes managerial conclusions	Material disclosure plus explicit human verification
Synthetic data or content	AI generates observations, scenarios, images, cases	Label clearly and preserve provenance
Autonomous workflow	Agent performs multi-step research or administrative operations	Full audit trail, access controls, and accountable owner

This is more informative than a yes/no field asking whether "AI was used."

Teaching and assessment

Watermarking creates a temptation to return to detection-based academic integrity. Universities should resist that temptation.

Turnitin itself tells instructors not to use its AI-writing score as the sole basis for adverse action because its generic detector can misidentify different forms of writing. Statistical

provider watermarks differ technically, yet the same procedural principle applies: institutions need corroboration, context, and human judgment. [9]

A university of economics can design assessments around demonstrable competence instead.

For quantitative courses, students can submit code, intermediate calculations, model diagnostics, variable definitions, sensitivity checks, and short oral explanations.

For marketing and management courses, students can document data sources, customer evidence, assumptions, alternatives considered, prompt or AI-use summaries where material, and managerial reasoning.

For economics courses, students can explain identification assumptions, causal mechanisms, robustness tests, model choice, and interpretation orally or in supervised settings.

For finance and accounting, students can defend valuation assumptions, reconcile source documents, explain formulas, and reproduce critical calculations.

Kenneth B.'s LinkedIn comment anticipates one consequence: increased use of oral assessment when written work becomes harder to attribute solely from the final artifact. 豰filecite孳turn0file0傢 A better approach is not to replace written work with oral exams wholesale. Universities can add targeted oral verification to assessments where authorship or competence is disputed.

Plagiarism and AI detection

The university should state explicitly in policy:

"AI watermark detected" does not mean "plagiarism detected."

Plagiarism involves attribution and use of others' work. A watermark concerns technical provenance or processing. An AI-generated passage can contain plagiarism, correctly attributed material, original synthesis, fabricated citations, or no borrowed material at all. Conversely, human-written text can be plagiarized without containing any AI mark.

Similarly:

"No watermark detected" does not mean "human-authored."

Detection coverage differs across vendors and modalities, marks can be absent from systems that do not implement them, transformations can affect signals, and short or low-information passages can provide less evidence for statistical detection. The watermarking literature itself treats detection as a statistical task rather than a universal identity test. [21]

A misconduct investigation should therefore use a hierarchy of evidence:

Course rules and disclosure requirements should come first. The institution should then consider source attribution, document history, data and code provenance, student explanation, prior work where relevant, and technical signals. A watermark result can contribute corroborating evidence but should not determine the decision alone.

Academic publishing

Watermarks may simplify some disclosure workflows if journals eventually incorporate reliable provenance verification. They may also produce new disputes when a manuscript was only edited, translated, or proofread by an AI system.

The safer publishing model is material-contribution disclosure. Authors should identify the AI system and explain what it did when its contribution falls within the journal's disclosure rules. Named humans retain responsibility for the manuscript. This closely matches the ICMJE model. [4]

University research offices should maintain a current matrix of major publisher and journal AI policies rather than imposing one wording across every discipline. Journal policies can change and disciplines apply different disclosure conventions.

Researchers should also preserve enough working material to answer an editor's question after submission: original notes, analysis code, reference-manager records, major draft versions, AI-use declarations, and validation steps. These records are stronger evidence of scholarly process than the final text alone.

Data sharing and research provenance

Watermarking has a distinct role in research-data and media provenance.

For synthetic datasets, generated survey responses, simulation outputs, AI-created cases, generated images, synthetic voices, or teaching videos, machine-readable provenance can provide meaningful metadata. C2PA's cryptographically signed manifest architecture is relevant because it gives systems a standardized way to associate provenance assertions with digital assets. [7]

Universities should avoid placing confidential prompts, personally identifiable information, unpublished hypotheses, peer-review material, or protected research data into provenance fields designed to travel with shared assets. Provenance design must respect data protection, research ethics, confidentiality, and intellectual-property rules.

For economic datasets, the university should emphasize conventional data provenance as strongly as AI provenance: source institution, download date, dataset edition, transformations, imputation, recoding, exclusion criteria, weighting, derivation of variables, and code used to create the analytical file.

Administrative policy and institutional communications

University communications present a stronger use case for visible disclosure than many student assignments. Synthetic images in recruitment, public campaigns, or news-style content can affect institutional trust. The communications office should therefore distinguish internally generated text assistance from synthetic media presented as documentary evidence.

For high-risk public-facing material, the university should preserve provenance and use clear human-readable labels when AI generation materially affects what audiences see or hear. This approach aligns with the EU's broader distinction between provider-side machine-readable marking and deployer-side disclosure obligations in relevant public-interest contexts. [22]

For routine administrative drafting, such as proofreading an email or restructuring a policy memo, mandatory visible labels on every document may create noise without adding useful information. A materiality threshold is more defensible.

Recommended institutional model

A university should adopt a "provenance plus responsibility" policy rather than an "AI detection" policy.

The governance model should contain six components.

First, define AI involvement by degree and function. Distinguish proofreading, translation, search assistance, coding assistance, substantive drafting, analytical interpretation, synthetic media generation, and autonomous task execution.

Second, require material disclosure rather than universal binary disclosure for human-authored academic work, while respecting journal, funder, legal, and assessment-specific requirements.

Third, prohibit automatic sanctions based solely on watermarking or generic AI-detection output. Technical evidence should enter an evidence review alongside policy, authorship records, source verification, drafts, code, and explanation. This is consistent with the human-judgment principle already stated by Turnitin for its own AI detector. [9]

Fourth, establish provenance standards for official synthetic media and research artifacts. Where compatible systems exist, preserve C2PA or equivalent Content Credentials rather than stripping metadata during institutional workflows. C2PA's technical model makes signed provenance a distinct asset-level capability. [7]

Fifth, redesign high-stakes assessment around demonstrated competence. Use oral defense, process artifacts, reproducible code, source logs, decision explanations, staged submissions, and supervised components according to learning objectives.

Sixth, build AI literacy around the difference between detection, provenance, attribution, authorship, plagiarism, truth, and accountability. These terms should appear separately in university policy. Treating them as synonyms creates both false accusations and false assurances.

Recommended decision rule

A practical university rule can be expressed as follows:

A watermark result may establish evidence of AI-system processing within the documented capabilities and error characteristics of the detector. It does not independently establish plagiarism, unauthorized assistance, authorship, factual unreliability, or academic misconduct.

This standard places the burden on the university to know what its technical evidence measures before drawing a disciplinary inference.

Stakeholder positions and consequences

This synthesis fits both the social-data evidence and the wider technical and policy record. Supporters correctly identify a growing need for trustworthy provenance. Critics correctly identify a limit: provenance signals become harmful when institutions treat them as authorship verdicts. 毅filecite孳turn0file0傢 The technical literature shows that machine-readable text watermarking is feasible at production scale. [21] European regulation gives such marking immediate governance relevance. [1] Academic publishing norms meanwhile continue to place responsibility, verification, attribution, and disclosure on identifiable humans. [4]

For a university of economics, this produces a clear strategic choice. Watermarking should become part of provenance infrastructure, not the foundation of

academic-integrity enforcement. The institution gains more by recording how research and student work were produced, requiring disclosure when AI materially contributes, teaching verification skills, preserving reproducible evidence, and assessing human understanding than by trying to classify every final paragraph as "human" or "AI."

The LinkedIn debate is therefore informative for university governance because its most consistent insight crosses both attitude clusters: knowing that AI touched a document is less important than knowing what AI contributed, what the human did, what evidence supports the work, and who accepts responsibility for the result. 毅filecite孳turn0file0傢

毅navlist孳Recent coverage of the Claude watermark debate孳turn19news34,turn19news37,turn19news32,turn19news35傢

---

[1] [11] Transparency obligations under Article 50 of the AI Act | Shaping Europe’s digital future

https://digital-strategy.ec.europa.eu/pt/node/17084?utm_source=chatgpt.com

[2] [15] A Watermark for Large Language Models

https://proceedings.mlr.press/v202/kirchenbauer23a.html?utm_source=chatgpt.com

[3] [7] Content Credentials : C2PA Technical Specification :: C2PA Specifications

https://spec.c2pa.org/specifications/specifications/2.2/specs/C2PA_Specification.html?utm_source=chatgpt.com

[4] ICMJE | Recommendations | Preparing a Manuscript for Submission to a Medical Journal

https://www.icmje.org/recommendations/browse/artificial-intelligence/ai-use-by-authors.html?utm_source=chatgpt.com

[5] Michael Wade - IMD business school for management and leadership courses

https://www.imd.org/faculty/professors/michael-wade?utm_source=chatgpt.com

[6] Code of Practice on Transparency of AI-generated Content | Shaping Europe’s digital future

https://digital-strategy.ec.europa.eu/en/policies/code-practice-ai-generated-content?utm_source=chatgpt.com

[8] [21] Scalable watermarking for identifying large language model outputs | Nature

https://www.nature.com/articles/s41586-024-08025-4?utm_source=chatgpt.com

[9] Using the AI Writing Report – Turnitin Guides

https://guides.turnitin.com/hc/en-us/articles/22774058814093-Using-the-AI-Writing-Report?utm_source=chatgpt.com

[10] [16] Advancing content provenance for a safer, more transparent AI ecosystem | OpenAI

https://openai.com/index/advancing-content-provenance/?utm_source=chatgpt.com

[12] [22] Signing the Code of Practice on Transparency of AI-generated Content | Shaping Europe’s digital future

https://digital-strategy.ec.europa.eu/en/faqs/signing-code-practice-transparency-ai-generated-content?utm_source=chatgpt.com

[13] Commission publishes second draft of Code of Practice on Marking and Labelling of AI-generated content | Shaping Europe’s digital future

https://digital-strategy.ec.europa.eu/en/library/commission-publishes-second-draft-code-practice-marking-and-labelling-ai-generated-content?utm_source=chatgpt.com

[14] Transparency obligations under Article 50 of the AI Act | Shaping Europe’s digital future

https://digital-strategy.ec.europa.eu/el/node/17084?utm_source=chatgpt.com

[17] Provenance signals (Content Credentials, SynthID) in OpenAI-generated content | OpenAI Help Center

https://help.openai.com/en/articles/8912793-c2pa-in-dall-e-3%23.woff2?utm_source=chatgpt.com

[18] Anthropic explains how Claude's invisible text watermarks will work

https://www.theverge.com/ai-artificial-intelligence/980869/anthropic-claude-watermarks-synthid-text-system?utm_source=chatgpt.com

[19] Beyond AI-Generated Labels: Watermarking, Co-Creation, and Conflation of AI-Generation with Disinformation

https://arxiv.org/abs/2607.13082?utm_source=chatgpt.com

[20] UNESCO survey: Two-thirds of higher education institutions have or are

https://www.unesco.org/en/articles/unesco-survey-two-thirds-higher-education-institutions-have-or-are-developing-guidance-ai-use?hub=103615&utm_source=chatgpt.com