CONTENT PROVENANCE · ASSISTIVE TECHNOLOGY · COMPELLED DISCLOSURE · DISABILITY DISCRIMINATION · DISPARATE IMPACT · UNIVERSAL DESIGN · PUBLIC ACCOMMODATION · SCREEN-OUT · DISCLOSURE PENALTY · NEW YORK CITY HUMAN RIGHTS LAW · PLATFORM GOVERNANCE · FORCED INTIMACY

The Human Signal

Reader-Facing AI Detection, the Disclosure Field, and the Tax on Assisted Authorship

> **A writer without a disability clears the badge by mentioning a habit. A writer who dictates through software clears it only by publishing a diagnosis.**

Travis Gilly, *The Human Signal: Reader-Facing AI Detection, the Disclosure Field, and the Tax on Assisted Authorship*, August 2026 (v0.5).

Travis Gilly
Real Safety AI Foundation

WHAT SHIPPED

A number beside your name, that any reader can summon

In July 2026 a publishing platform wired a commercial detector into the page readers actually see. Anyone reading a newly published post can run a scan and get back an estimate of how much of it a machine wrote, shown as a percentage next to the author's name.

No editor stands between the reader and the result. It is free, it is in the app, and it works on every eligible post. That is what separates this from a hidden watermark, which reports to specialists, and from a crowd flag, which reports to the platform.

> **What gets published alongside the writing, to exactly the same readership, is an allegation about it.**

SOURCES Substack, product announcement (July 2026); Perez, S., coverage of the integration.

THE SENTENCE THIS PAPER TAKES SERIOUSLY

Who counts as human when the writing runs through assistive technology?

> *“If we do not actively discriminate in favor of human content, then we're just gonna see more and more AI, and it's just gonna drown out any human signal that we have.”*

The detection company's chief executive, announcing a funding round the week after launch.

This paper takes that sentence at its word rather than treating it as a slip, and asks the question it leaves hanging.

SOURCES Bellan, R., reporting the chief executive's statement (2026).

AND THE REMEDY THAT CAME WITH IT

Beside the badge sits a box where you may explain yourself

The platform ships a field, controlled by the author, free text, shown to anyone who runs a scan, in which a writer can describe how the work gets made. Authors can also switch detection off for individual posts.

The design assumes that box is neutral: a channel open to everybody, costing each writer the same. That assumption is what this paper attacks.

SOURCES Substack, product announcement (July 2026).

PART ONE

Accuracy Is Not The Issue

the argument is strongest where the detector is right

I

An accurate instrument enforcing an incomplete taxonomy does not become fair. It becomes thorough.

THE ACCURACY GENERATION

Eight days later, a better model

The vendor shipped its next detector days after launch, claiming better than ninety-nine percent accuracy at spotting machine-assisted and mixed writing, and improved resistance to tools that lightly rewrite machine output to disguise it. Independent evaluation had already ranked it top of its category on the error that matters most to an accuser, being the only tool to hold a strict false-positive cap without losing accuracy.

At launch the platform itself allowed that nothing in this category is completely dependable, and that machine text somebody has edited is harder to place.

SOURCES Bellan, R. (2026); Press Release, Pangram Labs (Sept. 15, 2025); Bonifacic, I., launch coverage.

WHY THIS PAPER DECLINES THAT FIGHT

Assume it works perfectly. The problem is still there.

A classifier that reliably spots machine involvement will correctly spot it in the output of an assistive pipeline, because it really is there. The dictated draft was tidied by a model. The composed draft was generated at the writer's direction and revised by their own hand. Involvement is present, the instrument reports it, and the report is true.

Discrimination shows up one step further on, in how that label gets read, and in what different writers must pay to set the reading straight.

> **The taxonomy has two boxes, human and machine. Assisted authorship belongs in neither.**

PART TWO

What The Instrument Inherits

II

three findings that were already on the record

Readers cannot do this task, the instruments misread a predictable population, and explaining yourself costs money.

FIRST FINDING

Readers cannot do the job the badge does for them

This result is stable across the literature. People cannot reliably tell machine writing from human writing, and how confident they feel bears no measurable relation to whether they are right. In one survey most respondents were already familiar with machine text and still classified only fifteen percent of model passages correctly.

It is not only lay readers. Trained applied linguists could not identify who wrote research abstracts. Blinded physician reviewers rating clinical abstracts returned confidence figures statistically consistent with having no ability to tell at all.

> **Give somebody who cannot do the job a number that does it for them and you have not improved their judgement. You have only made them surer of it.**

SOURCES Ullah, U., et al. (2026); Mayor, E., Bietti, L. M., & Bangerter, A. (2025); Kendro, K., Maloney, J., & Jarvis, S. (2026); Lawrence, K. W., et al. (2024); Weber-Wulff, D., et al. (2023); Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023).

SECOND FINDING

The instruments misread a predictable population

60%+

the rate at which detectors classified essays by non-native English writers as machine-generated, while performing near-perfectly on native-speaker text

What trips these systems is orderly writing: grammar that holds steady, structure kept under control, sentences of similar weight, not much idiom. That description also fits a good deal of autistic writing, which tends to follow its rules and to mean what it says, and it fits what assistive pipelines put out generally, given that smoothing away the stumbles is the entire job.

SOURCES Liang, W., et al. (2023). GPT Detectors Are Biased Against Non-Native English Writers, 4 *Patterns* 100779.

THIRD FINDING

Saying you used a machine costs you, measurably

27,491

participants across sixteen preregistered experiments; ratings fell when evaluators believed a text was made by a model or with help from one

39.8%

drop in funds raised when a crowdfunding platform made disclosure mandatory, with backer numbers down 23.9 percent

The researchers call it the disclosure penalty. It runs through perceived authenticity, and it is sticky: it survived every metric, every context, every content type and every mitigation the literature had proposed. A separate programme of thirteen experiments found that anyone disclosing is trusted less, through diminished perceived legitimacy, and that this holds whether disclosure is voluntary or required.

SOURCES Raj, M., Berg, J. M., & Seamans, R. (2026). 155 *J. Experimental Psych.: Gen.* 896; Schilke, O., & Reimann, M. (2025). 188 *Org. Behav. & Hum. Decision Processes* 104405; Wang, N., & Liang, C. (2026). arXiv:2602.15698.

THE HONEST QUALIFICATION

And it sharpens the argument rather than softening it

A 2026 review of the journalism literature found no uniform penalty for news. Where the writing is routine and data-driven, provenance cues frequently produced no effect at all, or effects that depended on conditions.

So the penalty is sensitive to domain. Where the prose is impersonal and factual, it barely registers. Where the prose is imaginative, close, and spoken in somebody's own manner, it bites hardest. A paid newsletter is that second sort by definition, because voice is the product.

> **The instrument deploys its badge in exactly the register where the penalty it triggers is at its documented maximum.**

SOURCES Liçenji, L., & Hoxha, J. (2026). 9 *Frontiers A.I.* art. 1815243; Nakano, H., Takezawa, J., & Matulic, F. (2025). arXiv:2510.24011.

ONE MORE FINDING, WHICH PRICES THE BADGE

Being exposed costs more than admitting

The same trust programme found that although the disclosure penalty is robust, it is weaker than the effect of third-party exposure. Being revealed by somebody else damages trust more than revealing yourself does.

A badge any reader can summon is third-party exposure by design.

> **Which means the research has put the platform's two surfaces in order, and the order is bad news twice over for whoever is accused. Getting scanned hurts more than explaining, and explaining now hurts more than saying nothing used to.**

SOURCES Schilke, O., & Reimann, M. (2025), *supra*.

PART THREE

The Tax

the same box, priced differently for two populations

III

One writer answers with a preference. The other answers with a diagnosis.

THE BASELINE

Disability law already treats this information as yours

The employment rules bar asking an employee whether they are disabled, or how severe it is, unless the question is job-related and genuinely necessary, and wall off whatever is learned behind confidentiality. The public-accommodation rules run access without demanding paperwork even where fraud is the stated worry: staff meeting a service animal may ask two narrow questions and may not require documentation or ask about the nature or extent of somebody's disability.

So the system already knows the principle. Requiring somebody to hand over their diagnosis before they can take part like anyone else is the discrimination in itself, ahead of and separate from whatever they were trying to get to.

> **Disability scholarship has a name for the lived version. Forced intimacy: the permanent assumption that a disabled person will give up something private about themselves in order to receive what everyone else is simply handed.**

SOURCES 42 U.S.C. § 12112(d)(4)(A); 28 C.F.R. § 36.302(c)(6); Mingus, M., *Forced Intimacy: An Ableist Norm* (2017).

PRICE THE BOX

An adverse badge appears. What can each writer say?

The writer without a disability

Their truthful exculpatory answers are facts about craft. I write fast and clean. I use an editor. I dictate while walking and tidy it later, because I prefer to. None of that costs anything the law protects. It is the sort of thing writers volunteer in interviews.

The writer using assistive tools

Their truthful exculpatory answer is a medical fact. The machine signature is the accommodation; the dictation software and the model are there because of a condition. There is no version that clears the badge while keeping the condition back, because the condition is the explanation.

Answer vaguely, saying only that tools were involved, and you have accepted what the marker implies and taken the hit without touching the conclusion the reader who scanned has already drawn. Nothing but the adjustment explanation converts that machine fingerprint from something chosen into something required, and giving it means putting a diagnosis next to the writing, for good, in front of anybody who cares to look.

SOURCES This Article, Part IV.B; Raj, Berg, & Seamans (2026).

AND SILENCE STOPS BEING FREE

Once the box exists, saying nothing is an answer

Until that box appeared, keeping quiet about how you work was just keeping quiet about how you work. Build somewhere for the blameless to account for themselves and a blank space starts reading as a choice.

Nor does the per-post switch help. A writer whose posts alone cannot be scanned wears a different badge, and the reader completes it.

> **Better detection puts the price up rather than removing it, since getting things wrong was never what did the damage. What does the damage is a scheme of categories with nowhere to put assisted writing, applied by a machine that improves every quarter.**

PART FOUR

Why The Obvious Fixes Fail

a marked class is a disclosed class

IV

One sentence disposes of nearly every proposed repair.

THE INSTINCTIVE REPAIR

Add a third category, and you have built an announcement

The obvious fix is a new classification: let verified assisted writers carry an exemption tag instead of a percentage, or exempt their posts from scanning altogether. Every version of that recreates the harm it treats.

Put an assistive-technology marker on somebody and you have announced a disability while withholding which one, and withholding which one shields nothing worth shielding. The wound was never in the detail. It was in anybody finding out.

> **Category disclosure is disclosure.**

EVEN THE BEST-INTENTIONED VERSION

A carve-out that sorts disabled people by which disability they have

The European watermarking regime contains the most disability-aware provision in this field, exempting systems performing an assistive function for standard editing. But the line it draws runs straight through the middle of the disabled population.

It exempts communication aids that leave meaning untouched, while marking the composition aids that cognitive and executive-function disabilities depend on. A rule deciding which disabilities may pass unmarked is still a sorting of disabled people by machine-readable label.

SOURCES Regulation (EU) 2024/1689, art. 50(2); European Commission guidance construing the exception.

THE TEST

One sentence, and most proposals die on it

If a reader inspecting the badge, the tag, the empty box, the dispute record or the opt-out can infer anything at all about disability, the design has failed.

> **Put the proposed remedies through that filter and almost nothing gets out the other side.**

WHAT SURVIVES

Three designs that pass, and one thing they share

1 **Delete the badge, keep the author's own account**

No automatic percentage at all. The writer's own description is the only place provenance appears, so a disabled author's statement sits among statements rather than underneath a number arguing with it.

2 **One undifferentiated bucket**

If some indicator must exist, make it a single yes-or-no, tools were used, swallowing everything from a spellchecker to a full dictation pipeline. No gradations, no percentages. The assisted writer then stands invisibly inside the overwhelming majority of writers who touch some tool.

3 **A dispute path needing no reason**

Any correction or removal route open to every author, for any reason, with no reason shown and no documentation ever collected.

SOURCES This Article, Part V.B; Gedzyk-Nieman, S. A., & Derouin, A. (2025). 31 *Creative Nursing* 392.

WHY THEY SHARE IT

The disabled author disappears into a crowd

Herd anonymity is the accessibility property here. Granularity is the surveillance property. Universal design is not one repair among several in this setting. It is the only class of repair that does not itself disclose.

The accessibility literature arrives at the same place from the other direction, finding that relying only on adjustments people have to ask for slows help down, marks out whoever asks, and never reaches anyone who has not said anything. Which is why building access in from the start is treated as the correction rather than the supplement.

SOURCES Gedzyk-Nieman, S. A., & Derouin, A. (2025), *supra*.

PART FIVE

The Doctrine

two provisions, and a city statute that reaches the vendor

V

A standing system through which a platform administers the credibility of its authors.

TWO PROVISIONS

Screen-outs, and methods of administration

A business open to the public may not apply entry requirements that shut out disabled people, or tend to, from enjoying its services fully and equally, unless the requirement is genuinely needed. Separately it bars using standards, criteria or methods of administration whose effect is to discriminate on the basis of disability.

A detection-and-disclosure architecture is a method of administration in the plain sense of the phrase. What it amounts to is permanent operational machinery by which the platform manages how believable its writers appear.

SOURCES 42 U.S.C. § 12182(b)(2)(A)(i); 42 U.S.C. § 12182(b)(1)(D).

THE CLAIM, RESTATED

An accurate instrument enforcing an incomplete taxonomy

Three of these have now been built. First the hidden watermark, then the reader-reported authenticity flag, and now this, which finishes the job the other two began by turning up with a cure already bolted on.

That remedy is a box beside a number. For one population it costs a sentence about working habits. For the other it costs a diagnosis, published permanently, to every reader who cares to look. Both populations face the same box. Only one of them is being charged for it.

> **The harm was never the error rate. It is a taxonomy with no category for assisted authorship, enforced by an instrument getting better at enforcing it every month.**

SOURCES 42 U.S.C. § 12182(b)(1)(D); Raj, Berg, & Seamans (2026); Liang, W., et al. (2023).

REFERENCES

References

Travis Gilly, *The Human Signal: Reader-Facing AI Detection, the Disclosure Field, and the Tax on Assisted Authorship*, August 2026 (v0.5).

Americans with Disabilities Act of 1990, 42 U.S.C. §§ 12101–12213.

Bellan, R. (2026). As AI content floods the internet, Pangram raises $9M to detect it. *TechCrunch*. https://techcrunch.com/2026/07/29/as-ai-content-floods-the-internet-pangram-raises-9m-to-detect-it/.

Bonifacic, I. (2026). Substack is adding an AI detection feature. *Engadget*. https://www.engadget.com/2220064/substack-is-adding-an-ai-detection-feature/.

Carparts Distribution Center, Inc. v. Automotive Wholesaler's Association of New England, Inc., 37 F.3d 12 (1st Cir. 1994).

Coley, M. (2023). Guidance on AI detection and why we're disabling Turnitin's AI detector. *Vanderbilt University Brightspace*. https://www.vanderbilt.edu/brightspace/2023/08/16/guidance-on-ai-detection-and-why-were-disabling-turnitins-ai-detector/.

Doe v. Mutual of Omaha Insurance Co., 179 F.3d 557 (7th Cir. 1999).

Fair Housing Council of San Fernando Valley v. Roommates.com, LLC, 521 F.3d 1157 (9th Cir. 2008) (en banc).

Gedzyk-Nieman, S. A., & Derouin, A. (2025). Ensuring access for all nursing students: Classroom accommodations and universal design for learning. *Creative Nursing*, 31(4), 392–398. https://doi.org/10.1177/10784535251392561.

Kendro, K., Maloney, J., & Jarvis, S. (2026). Do large language models produce texts with "human-like" lexical diversity? Evidence from four ChatGPT models. *International Journal of Applied Linguistics*. https://doi.org/10.1111/ijal.70115.

Lawrence, K. W., Habibi, A. A., Ward, S. A., Lajam, C. M., Schwarzkopf, R., & Rozell, J. C. (2024). Human versus artificial intelligence-generated arthroplasty literature: A single-blinded analysis of perceived communication, quality, and authorship source. *The International Journal of Medical Robotics and Computer Assisted Surgery*, 20(1), e2621. https://doi.org/10.1002/rcs.2621.

Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, 19, art. 17. https://doi.org/10.1007/s40979-023-00140-5.

European Commission. (2026). Guidelines on the implementation of the transparency obligations for certain AI systems under Article 50 of the AI Act. https://digital-strategy.ec.europa.eu/en/library/guidelines-transparency-obligations-providers-and-deployers-ai-systems.

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, 4(7), 100779. https://doi.org/10.1016/j.patter.2023.100779.

Liçenji, L., & Hoxha, J. (2026). When news is "written by artificial intelligence": A systematic review of provenance and disclosure cues in journalism and their effects on credibility and trust. *Frontiers in Artificial Intelligence*, 9, art. 1815243. https://doi.org/10.3389/frai.2026.1815243.

Moloney, M., Brown, R. L., & Ciciurkaite, G. (2018). "Going the extra mile": Disclosure, accommodation, and stigma management among working women with disabilities. *Deviant Behavior*, 40(8), 942–956. https://doi.org/10.1080/01639625.2018.1445445.

Moody v. NetChoice, LLC, 603 U.S. 707 (2024).

Moreland, C. J., Pereira-Lima, K., Lee, K. T., et al. (2026). Disability accommodation access and requests in US internal medicine residents with disabilities. *JAMA Network Open*, 9(3), e263392. https://doi.org/10.1001/jamanetworkopen.2026.3392.

Nakano, H., Takezawa, J., & Matulic, F. (2025). Understanding reader perception shifts upon disclosure of AI authorship. *arXiv*. https://doi.org/10.48550/arxiv.2510.24011.

Mayor, E., Bietti, L. M., & Bangerter, A. (2025). Can large language models simulate spoken human conversations? *Cognitive Science*, 49(9), e70106. https://doi.org/10.1111/cogs.70106.

Mingus, M. (2017). Forced intimacy: An ableist norm. *Leaving Evidence*. https://leavingevidence.wordpress.com/2017/08/06/forced-intimacy-an-ableist-norm/.

New York City Human Rights Law, N.Y.C. Admin. Code §§ 8-101 to 8-131.

Pangram Labs. (2025). AI transparency company Pangram integrated into two leading news source and resource platforms [Press release]. *Business Wire*. https://www.businesswire.com/news/home/20250915315368/en.

Pereira-Lima, K., Meeks, L. M., Ross, K. E. T., et al. (2023). Barriers to disclosure of disability and request for accommodations among first-year resident physicians in the US. *JAMA Network Open*, 6(5), e239981. https://doi.org/10.1001/jamanetworkopen.2023.9981.

PGA Tour, Inc. v. Martin, 532 U.S. 661 (2001).

Raj, M., Berg, J. M., & Seamans, R. (2026). The artificial intelligence disclosure penalty: Humans persistently devalue AI-generated creative writing. *Journal of Experimental Psychology: General*, 155(4), 896–915. https://doi.org/10.1037/xge0001889.

Robles v. Domino's Pizza, LLC, 913 F.3d 898 (9th Cir. 2019).

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

Perez, S. (2026). Substack's new tool tells you who's been writing their newsletters with AI. *TechCrunch*. https://techcrunch.com/2026/07/22/substacks-new-tool-tells-you-whos-been-writing-their-newsletters-with-ai/.

Regulation (EU) 2024/1689 (Artificial Intelligence Act), art. 50(2), 2024 O.J. (L 1689).

Substack. (2026). Announcement of Pangram AI detection integration. *X*. https://x.com/Substack/status/2079598704424779787.

Communications Decency Act § 230, 47 U.S.C. § 230.

Ullah, U., Laudanna, S., Di Sorbo, A., Visaggio, C. A., & Murray, R. (2026). Breaking the imitation game: Can LLMs fool humans and machines alike? *International Journal of Intelligent Systems*, 2026(1), 8117975. https://doi.org/10.1155/int/8117975.

Wang, N., & Liang, C. (2026). How to disclose? Strategic AI disclosure in crowdfunding. *arXiv*. https://doi.org/10.48550/arxiv.2602.15698.

U.S. Department of Justice, Nondiscrimination on the Basis of Disability by Public Accommodations, 28 C.F.R. pt. 36.

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltynek, T., Guerrero-Dib, J., Popoola, O., Sigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19, art. 26. https://doi.org/10.1007/s40979-023-00146-z.

Weyer v. Twentieth Century Fox Film Corp., 198 F.3d 1104 (9th Cir. 2000).

Zhu, X., Jiang, F., & Volpone, S. D. (2025). The role of inclusive leadership in reducing disability accommodation request withholding. *Journal of Management*, 52(4), 1570–1601. https://doi.org/10.1177/01492063251314453.