AI TEXT DETECTION · ARTIFICIAL INTELLIGENCE DISCLOSURE PENALTY · ANNOTATOR BIAS · ALGORITHMIC AMPLIFICATION · CONTENT MODERATION · DISPARATE IMPACT · REASONABLE MODIFICATION · SURCHARGE DOCTRINE · PLATFORM ACCOUNTABILITY · SECTION 230 · SPEECH-TO-TEXT DICTATION · PROFESSIONAL NETWORKING PLATFORMS

# THE ONE-WAY VALVE

*Crowdsourced authenticity flagging, assistive technology, and disability discrimination on the professional platform*

On July 30, 2026, a major professional network gave every reader a button that accuses a post of being machine-made. The flag cuts the post's reach and trains the platform's own models. This deck is about who that button lands on, why the evidence says it cannot be fair, and the small procedural fix the law already supports.

Travis Gilly, *The One-Way Valve*, working draft v0.9 (August 2026)

**Travis Gilly** Real Safety AI Foundation

PART I

# THE MACHINE

*What shipped on July 30, and what it actually measures*

One new button, one hidden pipeline: readers accuse, the platform demotes, and the accusations become training data.

JULY 30, 2026

# A report button for “AI slop”

- Readers can flag any post as “AI slop,” meaning it feels machine-made to them.
- A flagged post travels less beyond the writer's own followers.
- The flags feed the platform's own models as training data.
- The company says no detector sits behind it; the judgment is human taste.

> *“Slop is hard to define and the definition changes; this lets us tune our models and make better feeds.”*
>
> LinkedIn Chief Product Officer Hari Srinivasan, July 31, 2026

**Sources** TechCrunch, *LinkedIn Adds a Button to Report AI-Generated ‘Slop’* (July 30, 2026), techcrunch.com/2026/07/30/linkedin-adds-a-button-to-report-ai-generated-slop/ · Fortune (July 31, 2026), fortune.com/2026/07/31/linkedin-seems-like-ai-slop-button-billions-automated-comments-attempts/ · Social Media Today (Aug. 2, 2026) · Inc. (Aug. 3, 2026)

THE DESIGN FLAW

# A valve that only turns one way

### WHAT THE BUTTON COLLECTS

An accusation: this reads as machine-made.

### WHAT IT NEVER COLLECTS

A “reads as human” vote from any reader.

The truth about who actually wrote the post.

Only one side of the judgment ever enters the system, and the truth never does. A model trained on that stream learns one thing: which writing gets complained about. Then the platform prices reach on it. Complaint-only channels also harvest the heavier emotion, because negative reactions are the ones people act on.

**Sources** Crawford & Gillespie, *What Is a Flag For? Social Media Reporting Tools and the Vocabulary of Complaint*, 18 New Media & Society 410 (2016) · Baumeister et al., *Bad Is Stronger than Good*, 5 Review of General Psychology 323 (2001) · Rozin & Royzman, *Negativity Bias, Negativity Dominance, and Contagion*, 5 Personality and Social Psychology Review 296 (2001)

THE ALTERNATIVE ALREADY EXISTS

# The gentler button was already there

The platform already has a button for readers who dislike a post. “Not interested” hides similar content from that reader's feed. It accuses nobody, and it trains nothing.

So everything a reader legitimately wants was already available. The only things the new button adds are the accusation, a penalty that follows the writer to other people's feeds, and the data pipeline built from the accusations.

In discrimination law, that matters: when a gentler tool already does the whole legitimate job, the harsher one has to answer for its extra harm.

**Sources** 42 U.S.C. § 12182(b)(2)(A)(ii); Payan v. L.A. Cmty. Coll. Dist., 11 F.4th 729, 738–40 (9th Cir. 2021)

A LIVE TEST NOBODY AUDITS

# Same writer, same process, four verdicts

**22% 39% 65% 71%**

Four posts, one unchanged writing process, four different “AI” scores from the scanner a publishing platform now shows readers. The tool's own “AI-assisted” category, the one built for openly assisted writing, read zero on all four, with the process stated right on each post. A stranger's casual, typo-filled post scored fully human. The vendor advertises one false accusation in ten thousand; independent testing of this tool class finds the low accusation rate is bought by defaulting doubtful text to “human.”

**Sources** Screen captures (Aug. 13, 2026) (on file with author) · Pangram Labs, *All About False Positives in AI Detectors* (Mar. 27, 2025), pangram.com/blog/all-about-false-positives-in-ai-detectors · Emi et al., *Technical Report on the Pangram AI-Generated Text Classifier*, arXiv:2402.14873 (2024) · Weber-Wulff et al., *Testing of Detection Tools for AI-Generated Text*, 19 International Journal for Educational Integrity, art. 26 (2023)

PART II

# THE RECORD

*Four bodies of evidence, one compounding failure*

People cannot sort machine text from human text reliably; the tools miss hardest at the margins; the crowd is a biased sensor; and disclosed help draws a penalty that is growing, not fading.

EVIDENCE LAYER ONE

# People cannot do this reliably

**49.9%**

evaluators sorting a newer model's text from human text, which is a coin flip

**70 / 62**

teachers and students picking the machine essay out of a pair

**15%**

machine text correctly caught when the model imitates a person's style

Confidence told you nothing; in the classroom study it was, in the researchers' words, essentially useless as a guide to accuracy, and neither chatbot experience nor subject expertise helped. Worst of all, the better a student's essay, the more it read as machine-made, and the “tells” people relied on were ordinary connectors like the word *overall*. The heuristic punishes disciplined writing as such.

**Sources** Clark et al., *All That's ‘Human’ Is Not Gold*, Proc. 59th Ann. Meeting of the Ass'n for Computational Linguistics 7282 (2021) · Waltzer et al., *Testing the Ability of Teachers and Students to Differentiate Between Essays Generated by ChatGPT and High School Students*, 2023 Human Behavior and Emerging Technologies, art. 1923981 · Ullah et al., *Breaking the Imitation Game*, 2026 International Journal of Intelligent Systems, art. 8117975

EVIDENCE LAYER TWO

# The tools miss hardest at the margins

**61.3% → 23.1%**

false accusations against non-native English writers: seven 2023 tools on average, then a 2026 rerun of the same essays on a current tool. Better, and still nearly one in four.

Ordinary polishing software alone trips the alarms, hitting non-native writers hardest. A sixty-thousand-post study found posts by likely-autistic writers flagged as machine-made significantly more often. And teams that engineered against the bias largely removed it, which means the bias is a choice, and a crowd with no engineering at all is the version of that choice this platform shipped.

**Sources** Liang et al., *GPT Detectors Are Biased Against Non-Native English Writers*, 4 Patterns 100779 (2023) · Al Ali, Helcl & Libovický, *Different Time, Different Language*, arXiv:2602.05769 (2026) · Bordalejo et al., 22 International Journal of Educational Technology in Higher Education, art. 6 (2025) · Chambers & Kelley, *The Misclassification of Autistic Writing as AI-Generated*, in AIED 2025 Proceedings, Part III (2025) · Jiang et al., Computers & Education (2024)

EVIDENCE LAYER THREE

# The crowd is a biased sensor, and the model scales it

- Ordinary raters flag text as abusive more often than trained experts reading the same words.
- Who the rater is moves the label: first language, age, and education each shift what identical text gets called.
- Posts in Black English were rated toxic at up to twice the rate, and models trained on those labels learned the habit and scaled it.
- The platform's stated plan is that member reports will tune its models, which is this same pipeline, pointed at writing style.

**Sources** Waseem, *Are You a Racist or Am I Seeing Things?*, Proc. First Workshop on NLP & Computational Social Science 138 (2016) · Al Kuwatly, Wich & Groh, Proc. Fourth Workshop on Online Abuse & Harms 184 (2020) · Sap et al., *The Risk of Racial Bias in Hate Speech Detection*, Proc. 57th Ann. Meeting of the Ass'n for Computational Linguistics 1668 (2019) · Xia, Field & Tsvetkov (2020) · Mozafari, Farahbakhsh & Crespi, 15 PLoS ONE e0237861 (2020) · Fortune (July 31, 2026)

EVIDENCE LAYER FOUR

# Disclosed help gets punished, and the mood is hardening

- People who use AI at work are rated less competent and less motivated, and the penalty follows them into hiring decisions.
- Across sixteen preregistered experiments and 27,491 people, written work lost value the moment readers believed a model wrote it or helped, and the penalty survived every fix the researchers tried.
- Disclosing AI use erodes trust whether the disclosure is voluntary or required, and even true, human-written headlines are believed less once labeled “AI.”
- The direction of travel is against assisted writers: communities adopting rules against AI content more than doubled in a single year, on stated grounds of quality and authenticity.

**Sources** Reif, Larrick & Soll, 122 Proceedings of the National Academy of Sciences e2426766122 (2025) · Raj et al., 155 Journal of Experimental Psychology: General 896 (2026) · Schilke & Reimann, Organizational Behavior and Human Decision Processes, art. 104405 (2025) · Altay & Gilardi, 3 PNAS Nexus pgae403 (2024) · Lloyd et al., Proc. 2025 CHI Conference on Human Factors in Computing Systems (2025)

PART III

# THE LAW

*The rule, the duty, and a remedy smaller than the fear*

Disability law reads effects, protects the assistive tool, supports a three-step procedural fix, and survives every version of the platform's legal shields.

THE RULE

# Effects count, and the tool is protected

Disability discrimination law does not require bad intent. It asks whether a neutral-looking practice ends up denying disabled people meaningful access to what everyone else gets, and it says a business may not run its systems in ways that screen them out.

> *The statute's own instruction: disability is judged "without regard to the ameliorative effects of mitigating measures such as ... assistive technology."*
>
> 42 U.S.C. § 12102(4)(E)(i)

The law tells decisionmakers to look past the assistive tool and see the person. The button does the reverse: it detects the tool's fingerprints in the writing and charges the writer for them.

**Sources** Alexander v. Choate, 469 U.S. 287, 299–301 (1985); Payan v. L.A. Cmty. Coll. Dist., 11 F.4th 729, 738–40 (9th Cir. 2021); Payan v. L.A. Cmty. Coll. Dist., 169 F.4th 971, 977–79 (9th Cir. 2026); 42 U.S.C. §§ 12102(4)(E)(i), 12182(b)(1)(D), (b)(2)(A)(i)

THE ASK

# The remedy is three small procedures

**1** — **A disclosure channel**

A writer can state, once, that their work is produced through assistive technology. No diagnosis required, nothing shown publicly unless the writer chooses.

**2** — **Notice**

When authenticity flags reduce a post's reach, the writer is told, in the analytics surface the platform has already said it is testing.

**3** — **Human review**

On appeal, a person decides whether the penalty stands, and flags on disclosed accounts stay out of model training while the review is pending.

None of this edits, removes, or compels anyone's speech. Every step is aimed at the platform's own penalty procedure.

**Sources** 42 U.S.C. § 12182(b)(2)(A)(ii); PGA Tour, Inc. v. Martin, 532 U.S. 661, 688 (2001)

THE SHIELD

# Section 230 does not block process

### THIRD CIRCUIT

The feed's curation is the platform's own speech, so the shield does not apply to claims aimed at the curation.

### NINTH CIRCUIT

Recommending content is publishing, so the shield holds; claims survive only by their structure.

### FIFTH CIRCUIT

Both shields can stack. The test: does the duty force the platform to monitor, alter, or remove other people's posts?

The three procedures force none of those, so the claim survives even the strictest circuit's test. And the Supreme Court settled the posture this year: the statute *"describes a defense, not an immunity,"* so a platform that invokes it litigates it to judgment.

**Sources** Anderson v. TikTok, Inc., 116 F.4th 180, 183–84 (3d Cir. 2024); Doe 1 v. Meta Platforms, Inc., No. 24-1672 (9th Cir. Apr. 28, 2026); Computer & Commc'ns Indus. Ass'n v. Paxton, No. 24-50721 (5th Cir. July 24, 2026); A.B. v. Salesforce, Inc., 123 F.4th 788, 794–95 (5th Cir. 2024); GEO Grp., Inc. v. Menocal, 607 U.S. 438, 449 (2026); Pers. Injury Plaintiffs v. TikTok, LLC, No. 24-7312 (9th Cir. Aug. 10, 2026); 47 U.S.C. § 230(c)(1)

THE PRICE

# You cannot sell the exit

> *The pricing rule: a business "may not impose a surcharge on a particular individual with a disability ... to cover the costs of measures" required for equal treatment.*
>
> 28 C.F.R. § 36.301(c)

The flag pipeline takes reach on the basis of writing style. Paid promotion sells reach. For everyone else, promotion buys extra visibility; for the flagged writer, the same purchase buys back the baseline others keep for free. Identical price list, discriminatory transaction. And the platform already sells promoted expert content as a product, so restoring an expert post's reach for money is its ordinary business, never a change to the service.

**Sources** 28 C.F.R. § 36.301(c); 28 C.F.R. § 35.130(f)

THE TRANSACTION

# Selling reach while throttling it

One claim never touches the shield at all. Courts enforce a specific promise a platform sells, and paid promotion is exactly that: money in exchange for distribution of one identified post. Selling that promise while the seller's own systems throttle the buyer's class of posts is a deceptive trade practice under federal and Illinois law.

Knowledge is not a hurdle. The company publicly counts what it controls: hundreds of thousands of blocked fake comments every day, and billions of blocked automation attempts in a couple of months. A seller who measures the integrity of distribution that precisely knows exactly what it is selling.

**Sources** Barnes v. Yahoo!, Inc., 570 F.3d 1096, 1107–09 (9th Cir. 2009); Doe v. Grindr Inc., 128 F.4th 1148 (9th Cir. 2025); Est. of Bride v. Yolo Techs., Inc., 112 F.4th 1168, 1173, 1178–79 (9th Cir. 2024); 15 U.S.C. § 45(a); 815 ILCS 505/2; Fortune (July 31, 2026)

THE PROOF IT IS FEASIBLE

# Europe already requires all of it

- A written reason every time a post's visibility is restricted, demotion included.
- A complaint system where the decision is never made by machine alone; a human is in the loop.
- Every one of those decisions filed to a public database anyone can audit.

The company already runs this process today for its European users, because European law requires it. So the burden argument collapses into one sentence: the only question left is whether American disability law entitles American disabled writers to the process the platform already operates for everyone in Europe.

**Sources** Regulation (EU) 2022/2065 (Digital Services Act), arts. 17, 20, 24(5), 2022 O.J. (L 277) 1

THE CLAIM

# The distance is the case

The fair version of this feature has existed on the platform for years, and it still runs. The July 30 button added only three things: the accusation, the penalty, and the machine that learns from the accusations. That distance, between the tool the platform already had and the one it shipped, is the measure of the case.

What should happen next is small: a disclosure channel, notice, and a human on the appeal. The law does not ask the platform to referee authenticity. It asks for process before penalty, and it forbids selling the exit.

**Sources** Travis Gilly, *The One-Way Valve*, working draft v0.9 (August 2026), Parts IV–X

## CASES

Travis Gilly, *The One-Way Valve: Crowdsourced Authenticity Flagging, Assistive Technology, and Disability Discrimination on the Professional Platform*, working draft v0.9 (August 2026). Every entry appears exactly as the paper lists it.

A.B. v. Salesforce, Inc., 123 F.4th 788 (5th Cir. 2024).

Alexander v. Choate, 469 U.S. 287 (1985).

Alexander v. Sandoval, 532 U.S. 275 (2001).

Anderson v. TikTok, Inc., 116 F.4th 180 (3d Cir. 2024).

Barnes v. Yahoo!, Inc., 570 F.3d 1096 (9th Cir. 2009).

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

Computer & Communications Industry Association v. Paxton, No. 24-50721 (5th Cir. July 24, 2026).

Cummings v. Premier Rehab Keller, P.L.L.C., 596 U.S. 212 (2022).

Doe 1 v. Meta Platforms, Inc., No. 24-1672 (9th Cir. Apr. 28, 2026).

Doe v. BlueCross BlueShield of Tennessee, Inc., 926 F.3d 235 (6th Cir. 2019).

Doe v. Grindr Inc., 128 F.4th 1148 (9th Cir. 2025).

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

Estate of Bride v. Yolo Technologies, Inc., 112 F.4th 1168 (9th Cir. 2024).

Ford v. Schering-Plough Corp., 145 F.3d 601 (3d Cir. 1998).

GEO Group, Inc. v. Menocal, 607 U.S. 438 (2026).

Gonzalez v. Google LLC, 598 U.S. 617 (2023) (per curiam).

HomeAway.com, Inc. v. City of Santa Monica, 918 F.3d 676 (9th Cir. 2019).

In re Apple Inc. App Store Simulated Casino-Style Games Litigation, 625 F. Supp. 3d 971 (N.D. Cal. 2022).

Jacobson v. Delta Airlines, Inc., 742 F.2d 1202 (9th Cir. 1984).

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

National Association of the Deaf v. Harvard University, 377 F. Supp. 3d 49 (D. Mass. 2019).

NCAA v. Smith, 525 U.S. 459 (1999).

Parker v. Metropolitan Life Insurance Co., 121 F.3d 1006 (6th Cir. 1997) (en banc).

Payan v. Los Angeles Community College District, 11 F.4th 729 (9th Cir. 2021).

Payan v. Los Angeles Community College District, 169 F.4th 971 (9th Cir. 2026).

Personal Injury Plaintiffs v. TikTok, LLC, No. 24-7312 (9th Cir. Aug. 10, 2026).

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

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

Sikhs for Justice, Inc. v. Facebook, Inc., 144 F. Supp. 3d 1088 (N.D. Cal. 2015).

U.S. Department of Transportation v. Paralyzed Veterans of America, 477 U.S. 597 (1986).

Wetzel v. Glen St. Andrew Living Community, LLC, 901 F.3d 856 (7th Cir. 2018).

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

## STATUTES AND REGULATIONS

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

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

Federal Grant and Cooperative Agreement Act, 31 U.S.C. §§ 6301–6309.

Federal Trade Commission Act § 5, 15 U.S.C. § 45.

Illinois Consumer Fraud and Deceptive Business Practices Act, 815 ILCS 505/1 et seq.

Nondiscrimination on the Basis of Disability by Public Accommodations, 28 C.F.R. § 36.301.

Nondiscrimination on the Basis of Disability in State and Local Government Services, 28 C.F.R. § 35.130.

Regulation (EU) 2022/2065 (Digital Services Act), 2022 O.J. (L 277) 1.

Rehabilitation Act of 1973 § 504, 29 U.S.C. § 794.

## SECONDARY AUTHORITIES

Al Ali, A., Helcl, J., & Libovický, J. (2026). Different time, different language: Revisiting the bias against non-native speakers in GPT detectors (arXiv:2602.05769).

Al Kuwatly, H., Wich, M., & Groh, G. (2020). Identifying and measuring annotator bias based on annotators' demographic characteristics. *Proceedings of the Fourth Workshop on Online Abuse and Harms*, 184.

Altay, S., & Gilardi, F. (2024). People are skeptical of headlines labeled as AI-generated, even if true or human-made, because they assume full AI automation. *PNAS Nexus*, 3(10), pgae403.

Anthropic Help Center. (2026). *Text watermarking on Claude models*. https://support.claude.com/en/articles/16266773.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, 5, 323.

Bordalejo, B., et al. (2025). “Scarlet Cloak and the Forest Adventure”: A preliminary study of the impact of AI on commonly used writing tools. *International Journal of Educational Technology in Higher Education*, 22(1), art. 6.

Chambers, S., & Kelley, M. C. (2025). The misclassification of autistic writing as AI-generated. In A. I. Cristea et al. (Eds.), *Artificial intelligence in education: 26th international conference, AIED 2025, proceedings, part III*. Springer.

Clark, E., et al. (2021). All that's “human” is not gold: Evaluating human evaluation of generated text. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 7282.

Crawford, K., & Gillespie, T. (2016). What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media & Society*, 18, 410.

Emi, B., et al. (2024). *Technical report on the Pangram AI-generated text classifier* (arXiv:2402.14873). Pangram Labs.

Haupt, M., Freidank, J., & Haas, A. (2024). Consumer responses to human-AI collaboration at organizational frontlines: Strategies to escape algorithm aversion in content creation. *Review of Managerial Science*, 19(2), 377–413.

Jiang, Y., et al. (2024). Detecting ChatGPT-generated essays in a large-scale writing assessment: Is there a bias against non-native English speakers? *Computers & Education*, 224.

Liang, W., et al. (2023). GPT detectors are biased against non-native English writers. *Patterns*, 4, 100779.

Lloyd, T., et al. (2025). AI rules? Characterizing Reddit community policies towards AI-generated content. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*.

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2020). Hate speech detection and racial bias mitigation in social media based on BERT model. *PLoS ONE*, 15, e0237861.

Nakano, H., et al. (2025). Understanding reader perception shifts upon disclosure of AI authorship. *Proceedings of the 31st International Conference on Intelligent User Interfaces*.

Prajod, P., Cools, H., & Röggla, T. (2026). Full disclosure, less trust? How the level of detail about AI use in news writing affects readers' trust (arXiv:2601.09620).

Raj, M., et al. (2026). The artificial intelligence disclosure penalty: Humans persistently devalue AI-generated creative writing. *Journal of Experimental Psychology: General*, 155(4), 896–915.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, 5, 296.

Reif, J. A., Larrick, R. P., & Soll, J. B. (2025). Evidence of a social evaluation penalty for using AI. *Proceedings of the National Academy of Sciences*, 122(19), e2426766122.

Sap, M., et al. (2019). The risk of racial bias in hate speech detection. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1668.

Schilke, O., & Reimann, M. (2025). The transparency dilemma: How AI disclosure erodes trust. *Organizational Behavior and Human Decision Processes*, article 104405.

Toff, B., & Simon, F. M. (2024). “Or they could just not use it?”: The dilemma of AI disclosure for audience trust in news. *The International Journal of Press/Politics*, 30(4), 881–903.

Ullah, U., Laudanna, S., Di Sorbo, A., & Visaggio, C. A. (2026). Breaking the imitation game: Can LLMs fool humans and machines alike? *International Journal of Intelligent Systems*, 2026, art. 8117975.

Waltzer, T., Cox, R. L., & Heyman, G. D. (2023). Testing the ability of teachers and students to differentiate between essays generated by ChatGPT and high school students. *Human Behavior and Emerging Technologies*, 2023, art. 1923981.

Waseem, Z. (2016). Are you a racist or am I seeing things? Annotator influence on hate speech detection on Twitter. *Proceedings of the First Workshop on NLP and Computational Social Science*, 138.

Weber-Wulff, D., et al. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19.

Wich, M., Al Kuwatly, H., & Groh, G. (2020). Investigating annotator bias with a graph-based approach. *Proceedings of the Fourth Workshop on Online Abuse and Harms*, 191.

Xia, M., Field, A., & Tsvetkov, Y. (2020). Demoting racial bias in hate speech detection. *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, 7.

Zhang, Y., & Gosline, R. (2023). Human favoritism, not AI aversion: People's perceptions (and bias) toward generative AI, human experts, and human–GAI collaboration in persuasive content generation. *Judgment and Decision Making*, 18, e41.