The One-Way Valve:

Crowdsourced Authenticity Flagging, Assistive Technology, and Disability Discrimination on the Professional Platform

Travis Gilly* August 2026 (v0.9)

Abstract

On July 30, 2026, LinkedIn shipped a report option labeled "Seems like AI slop," inviting members to flag posts that read as machine-made, reducing the distribution of flagged posts, and feeding the flags into the platform's classifiers as training signal. This Article identifies the instrument's defining defect: it is a one-way valve. It collects accusations and nothing else. No countervailing judgment enters, ground truth never enters, and the resulting classifier learns complaint prevalence rather than authorship. The empirical record supplies each layer of the resulting injury. Humans cannot reliably distinguish machine text from human text, and their confidence bears no relation to their accuracy; crowd raters carry measured demographic biases that trained models absorb and amplify; evaluators penalize disclosed AI assistance even where the work is accurate and the author is human; and the writers most likely to be flagged write in the disciplined registers associated with disability, second-language acquisition, and assistive technology. The Article then maps the doctrine: disparate impact and meaningful access under the Americans with Disabilities Act and Section 504, the surcharge prohibition as applied to paid reach restoration, a procedural modification remedy that requires no alteration of any third party's content, and the resulting path through Section 230 in every circuit configuration, including the Fifth Circuit's July 2026 holding that the statutory shield and the First Amendment shield stack.

Keywords: 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.

---

*Travis Gilly is Founder and Executive Director of the Real Safety AI Foundation. ORCID: 0009-0007-2954-6313. Correspondence: t.gilly@ai-literacy-labs.org. The author discloses the use of speech-to-text software (Wispr Flow) and Anthropic's Claude as assistive technology for a diagnosed neurodevelopmental condition. All legal analysis, case synthesis, theoretical framing, doctrinal arguments, and conclusions are solely the author's own.

Contents

I. Introduction

On July 30, 2026, LinkedIn added a new option to its reporting menu: "Seems like AI slop."[1] The company's Chief Product Officer described the feature as a channel for human taste rather than a detector: members report what reads to them as machine-made, the platform reduces the distribution of what they report, and the reports become training signal for LinkedIn's own models.[2] Three days later, Anthropic began watermarking the text output of its newest models worldwide.[3] Nine days before the button shipped, Substack integrated a commercial AI-text classifier and began displaying its verdicts, including to readers, on any eligible post.[4]

The infrastructure of authenticity judgment arrived in one month, on three platforms, in three forms: a crowd vote, a hidden mark, and a vendor score. This Article takes the crowd vote as its subject because it is the most legally consequential of the three and the least examined. The literatures on detector error and on academic-integrity enforcement are developed; the crowdsourced flag, deployed against professional speech outside any educational setting, wired to distribution, and used as training data, has no treatment in the legal literature.

The test case that organizes the analysis is the disabled professional who writes through disclosed assistive technology. Speech-to-text systems route dictated speech through a language model that punctuates, formats, and repairs the transcript; drafting assistants render dictated

---

1. TechCrunch, *LinkedIn Adds a Button to Report AI-Generated 'Slop'* (July 30, 2026), `https://techcrunch.com/2026/07/30/linkedin-adds-a-button-to-report-ai-generated-slop/` (last visited Aug. 13, 2026).

2. Hari Srinivasan, quoted in *LinkedIn Adds a 'Seems Like AI Slop' Button After Blocking Billions of Automated Comment Attempts*, Fortune (July 31, 2026), `https://fortune.com/2026/07/31/linkedin-seems-like-ai-slop-button-billions-automated-comments-attempts/` (last visited Aug. 13, 2026) ("Slop is hard to define and the definition changes; this lets us tune our models and make better feeds."); Chris Morris, *LinkedIn Just Added a Button to Report AI Slop*, Inc. (Aug. 3, 2026), `https://www.inc.com/chris-morris/linkedin-just-added-a-button-to-report-ai-slop-theres-just-1-problem/91383248` (last visited Aug. 13, 2026) (reporting that member feedback determines the reach a post receives outside the poster's network).

3. Anthropic Help Ctr., *Text Watermarking on Claude Models*, `https://support.claude.com/en/articles/16266773` (last visited Aug. 13, 2026) (describing an imperceptible statistical watermark applied to text from models launched on or after August 2, 2026, and cautioning that the presence of the mark does not establish authorship where a model was used for proofreading, translation, or similar assistance).

4. Screen captures of the Substack and Pangram integration, including the eligibility notice stating that posts published before July 21, 2026, 10:30 AM CDT are not scanned, and a scan panel rendered on a third party's post (Aug. 13, 2026) (on file with author).

argument into finished prose. The thoughts, claims, and conclusions are the writer's; the surface of the text carries the machine's fingerprints, because a machine touched the surface. That writer publishes under a disclosure of exactly this process. The question the new instrument poses is what happens to that writer when the surface of the text is put to a vote.

The Article makes four claims. First, as a matter of instrument design, the button is a one-way valve: it collects only accusations, admits no countervailing signal and no ground truth, and therefore trains the platform's classifier on complaint prevalence rather than on authorship. Second, the empirical record establishes every link in the resulting causal chain: humans cannot perform the discrimination task the button assigns; the crowd performing it carries measured demographic biases that models trained on crowd labels absorb and amplify; evaluators penalize disclosed AI assistance even when the work is accurate and human; and the prose registers most likely to draw the flag are the registers of disability, second-language acquisition, and assistive technology. Third, disability law reaches the design: the flag pipeline screens out and denies meaningful access on the basis of disability under Title III of the Americans with Disabilities Act and, through the platform's corporate parent, Section 504 of the Rehabilitation Act, and paid reach restoration prices the accommodation in violation of the surcharge prohibition. Fourth, the remedy is modest and survives Section 230 in every circuit configuration, because it is procedural: a disclosure channel, notice, and an appeal, none of which requires the platform to monitor, alter, or remove any third party's content.

Part II describes the mechanism as shipped and formalizes the one-way valve. Part III assembles the empirical record in four layers. Parts IV through VII develop the disability doctrine: disparate impact and meaningful access, Title III coverage and the procedural modification, the surcharge, and Section 504. Part VIII maps Section 230 and the First Amendment across the Third, Ninth, and Fifth Circuits. Part IX develops a deception theory around paid distribution that sits outside Section 230 entirely. Part X measures the design against the European Union's Digital Services Act, which prohibits most of it expressly.

II. The Mechanism

A. The Button as Shipped

The July 30 announcement supplies the operative facts, and they are best read as a set of admissions. The reporting option appears in the standard flag menu on feed posts. A member who selects it hides the post from that member's own feed. Reports reduce the flagged post's distribution beyond the author's follower graph, a treatment the company analogized to its existing "not interested" signal.[5] Aggregated reports feed the platform's content classifiers as training data.[6] The company was explicit that no automated detector produces the judgment; the design solicits member reports because, in its Chief Product Officer's words, slop "is hard to define and the definition changes," and the reports "let[] us tune our models."[7] A companion change replaced the composer's generative drafting assistant with a proofreading assistant.[8] The company has separately announced, and as of this writing has not shipped, an analytics notification that would tell authors their post has been flagged.[9]

B. A One-Way Instrument

Strip the branding and describe the instrument. It solicits a binary judgment from untrained raters on a question of fact: whether a machine produced this text. It records only one of the two possible answers. A reader who finds a post evidently human has no button to press; the judgment "reads as human" is never collected, never weighed against the accusation, and never enters the training corpus. No ground truth enters at any point, because the platform does not know, and disclaims knowing, which posts are in fact machine-made. The label distribution

---

5. *Supra* note 2.

6. *Id.*

7. *Supra* note 2.

8. *Supra* note 1.

9. Andrew Hutchinson, *LinkedIn Offers the Option to Report AI Slop*, Soc. Media Today (Aug. 2, 2026), `https://www.socialmediatoday.com/news/linkedin-offers-the-option-to-report-ai-slop/826781/` (last visited Aug. 13, 2026) (quoting Srinivasan: "For anyone who shares content, we will test a way to privately flag, in your analytics dashboard, when members feel your post may have come off as inauthentic or heavy use of AI").

the classifier trains on is therefore not authorship and not even the crowd's full judgment of authorship. It is complaint prevalence: the rate at which each post provokes a specific hostile reaction from the subset of readers moved to act on it. The defect is not novel to this platform; the sociology of the report button has long described the flag as a vocabulary of complaint rather than an instrument of measurement, and the psychology of evaluation adds that negative information dominates positive across domains, so an accusation-only channel harvests the heavier tail by design.[10]

This is selection on the dependent variable, built into the sensor. Any classifier trained on the output learns the features of prose that attract accusation, and the platform then prices distribution on those features. The company's own framing completes the point. By disclaiming any detector and soliciting human feeling, the platform concedes that the instrument measures perception. Perception is exactly the quantity the empirical record in Part III shows to be unreliable, demographically skewed, and animus-laden. The design does not fail to filter those defects; it has no component whose job is to filter them, and it has a component whose job is to scale them.

C. The Redundant Alternative

The platform already operates a fully adequate instrument for the interest it invokes. The "not interested" control predates the button, reduces similar content in the reporting member's feed, and carries no accusation of fraud and no authenticity training signal. Every legitimate reader interest the slop button serves, the existing control already served. What the new button adds is only the accusatory layer: the authenticity verdict, the cross-member distribution penalty, and the classifier trained on the accusations. When the analysis reaches the modification and disparate

10. Kate Crawford & Tarleton Gillespie, *What Is a Flag For? Social Media Reporting Tools and the Vocabulary of Complaint*, 18 New Media & Soc'y 410 (2016); Roy F. Baumeister et al., *Bad Is Stronger than Good*, 5 Rev. Gen. Psychol. 323 (2001); Paul Rozin & Edward B. Royzman, *Negativity Bias, Negativity Dominance, and Contagion*, 5 Personality & Soc. Psychol. Rev. 296 (2001).

impact doctrine in Parts IV and V, this fact does the work of the less discriminatory alternative, and the platform built it itself.[11]

D. Adjacent Instruments

Two contemporaneous deployments frame the button and supply its evidentiary context.

The first is the watermark. Anthropic's mark is imperceptible, statistical, and applied at generation; the company's own documentation cautions that the mark indicates a model's involvement in producing text, and does not establish that the ideas, argument, or authorship are the model's, expressly instancing proofreading and translation.[12] The vendor that built the strongest provenance signal in the ecosystem disclaims precisely the inference the crowd is being asked to vote on.

The second is the vendor score. Substack's integrated classifier is marketed on a false positive rate of one in ten thousand.[13] That figure describes one error direction only: human text accused of being machine-made. The independent testing literature finds that detection tools purchase low accusation rates by defaulting borderline text toward the human label, a bias documented across the tool class.[14] The author's own publication record, produced through the disclosed pipeline described in the opening footnote and constant across posts, supplies a deployment test: four posts scanned by the integrated classifier returned 22, 39, 65, and 71 percent "AI," two of the four carried the top-line label "Mostly Human-Written," and all four returned zero percent in the classifier's own "AI-assisted" category, against an authorial

---

11. *See* 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).

12. *Supra* note 3.

13. Pangram Labs, *All About False Positives in AI Detectors* (Mar. 27, 2025), `https://www.pangram.com/blog/all-about-false-positives-in-ai-detectors` (last visited Aug. 13, 2026) (claiming an "industry-leading 1 in 10,000 false positive rate"); *see also* Bradley Emi et al., *Technical Report on the Pangram AI-Generated Text Classifier*, arXiv:2402.14873 (2024) (claiming error rates far below competing detectors and no bias against nonnative English speakers).

14. Debora Weber-Wulff et al., *Testing of Detection Tools for AI-Generated Text*, 19 Int'l J. for Educ. Integrity, art. 26 (2023) (twelve public tools and two commercial systems; concluding the tools are neither accurate nor reliable and exhibit a main bias toward classifying output as human-written).

disclosure of AI assistance pinned to each post.[15] A constant authorship process returned a forty-nine point spread. The score tracks the surface register of individual posts; the surface register is the thing the score is marketed as seeing through. The instrument class cannot see the assisted category that names the accommodation, and the crowd instrument examined here was designed with no component that could.

III. The Empirical Record

A. Humans Cannot Perform the Task Reliably

The button assigns readers a discrimination task. The task has been measured. Evaluators presented with model-generated and human text performed at 57.9 percent on an older model's output and at chance, 49.9 percent, on the successor's.[16] High school teachers reading paired essays, one written by a student and one by a chatbot, identified the machine at 70 percent; their students managed 62.[17] Two findings in that study matter more than the accuracy figure. Self-reported confidence was, in the authors' words, so poor an indicator of accuracy as to be essentially useless, and neither prior experience with the chatbot nor subject-matter expertise predicted performance.[18] And the student essays judged higher in quality were the ones hardest to distinguish from the machine's.[19] The heuristic the crowd actually uses penalizes disciplined prose as such. Where a model is instructed to imitate a particular human style, performance collapses further: in one 2026 study, participants correctly identified fifteen percent of the machine text.[20]

---

15. Screen captures (Aug. 13, 2026) (on file with author). A contemporaneous casual post by a third author, informal and carrying uncorrected typographical errors, returned "Fully Human-Written" at one hundred percent. *Id.*

16. Elizabeth Clark et al., *All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text*, in PROC. 59TH ANN. MEETING OF THE ASS'N FOR COMPUTATIONAL LINGUISTICS 7282 (2021).

17. Tal Waltzer et al., *Testing the Ability of Teachers and Students to Differentiate Between Essays Generated by ChatGPT and High School Students*, 2023 Hum. Behav. & Emerging Techs., art. 1923981.

18. *Id.*

19. *Id.* Participating teachers reported treating transitional connectives and the word "overall" as machine tells.

20. Ubaid Ullah et al., *Breaking the Imitation Game: Can LLMs Fool Humans and Machines Alike?*, 2026 Int'l J. Intelligent Sys., art. 8117975.

B. The Writers the Instruments Misread

Where the task is delegated to machines, the errors concentrate on marginalized writers. In the 2023 detector generation, seven widely used tools misclassified essays by nonnative English speakers as machine-generated at an average rate of 61.3 percent, and 97.8 percent of those essays were flagged by at least one detector, while the same tools read native speakers' essays as human.[21] A 2026 follow-up reran the question and reported the current state in both directions: no systematic bias against nonnative writers of Czech, and, on Liang's own English corpus, a contemporary commercial detector whose nonnative false positive rate had fallen to 23.1 percent, a reduction of nearly two thirds that still leaves roughly one nonnative essay in four falsely accused.[22] Assistive polishing tools of the kind LinkedIn now offers inside its own composer trigger the same detectors: text merely improved for readability by grammar software drew false positives, with nonnative writers again bearing the concentration.[23] Peer-reviewed work extends the pattern to autistic writers: in a corpus of roughly sixty thousand social media posts split into likely-autistic and general subcorpora, texts from the likely-autistic subcorpus were flagged as machine-generated at significantly elevated rates, though overall flag rates in both subcorpora were low, and the authors call for ethical scrutiny of the detection models themselves.[24]

The counter-literature is instructive rather than exculpatory. Purpose-built detectors trained with care have eliminated the nonnative bias on controlled corpora,[25] and the vendor whose classifier Substack deployed reports no such bias in its own evaluation.[26] Bias in this instrument class is an engineering choice. That is the point. A platform choosing among

---

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

22. Adnan Al Ali, Jindřich Helcl & Jindřich Libovický, *Different Time, Different Language: Revisiting the Bias Against Non-Native Speakers in GPT Detectors*, arXiv:2602.05769 (2026).

23. Barbara Bordalejo et al., *"Scarlet Cloak and the Forest Adventure": A Preliminary Study of the Impact of AI on Commonly Used Writing Tools*, 22 Int'l J. Educ. Tech. Higher Educ., art. 6 (2025).

24. Summer Chambers & Matthew C. Kelley, *The Misclassification of Autistic Writing as AI-Generated*, in ARTIFICIAL INTELLIGENCE IN EDUCATION: 26TH INT'L CONFERENCE, AIED 2025, PROCEEDINGS, PART III (Alexandra I. Cristea et al. eds., 2025).

25. Yang Jiang et al., *Detecting ChatGPT-Generated Essays in a Large-Scale Writing Assessment: Is There a Bias Against Non-Native English Speakers?*, Computers & Educ. (2024).

26. *Supra* note 13.

instruments chose the one with no engineering at all: an untrained crowd, voting on register, wired to a classifier.

C. The Crowd as Biased Instrument, and the Amplifier Behind It

The crowd is not a neutral sensor. Non-expert raters label text as abusive at higher rates than experts asked the same question.[27] Raters' own demographics drive the labels: first language, age, and education each produced statistically significant differences in what identical text was called.[28] In the most studied instance, crowd annotators rated posts in African American English as toxic at inflated rates, and models trained on those labels acquired and propagated the bias, flagging AAE posts at up to twice the rate of others; telling raters the dialect of what they were reading significantly reduced the effect.[29] Successor work confirmed the amplification step across architectures: classifiers trained on crowd labels absorb the raters' skew and scale it.[30]

That is the pipeline the platform announced, by name. Member reports tune the models.[31] The literature above spent seven years documenting what that pipeline does to the speech registers of protected classes when the labels come from an untrained crowd. The platform rebuilt it, aimed it at prose register, and attached it to distribution.

---

27. Zeerak Waseem, *Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter*, in PROC. FIRST WORKSHOP ON NLP & COMPUTATIONAL SOC. SCI. 138 (2016) (finding characterized consistently in Al Kuwatly et al. and Xia et al., *infra*).

28. Hala Al Kuwatly, Maximilian Wich & Georg Groh, *Identifying and Measuring Annotator Bias Based on Annotators' Demographic Characteristics*, in PROC. FOURTH WORKSHOP ON ONLINE ABUSE & HARMS 184 (2020).

29. Maarten Sap et al., *The Risk of Racial Bias in Hate Speech Detection*, in PROC. 57TH ANN. MEETING OF THE ASS'N FOR COMPUTATIONAL LINGUISTICS 1668 (2019).

30. Mengzhou Xia, Anjalie Field & Yulia Tsvetkov, *Demoting Racial Bias in Hate Speech Detection*, in PROC. EIGHTH INT'L WORKSHOP ON NAT. LANGUAGE PROCESSING FOR SOC. MEDIA 7 (2020); Marzieh Mozafari, Reza Farahbakhsh & Noël Crespi, *Hate Speech Detection and Racial Bias Mitigation in Social Media Based on BERT Model*, 15 PLoS ONE e0237861 (2020); *see also* Maximilian Wich, Hala Al Kuwatly & Georg Groh, *Investigating Annotator Bias with a Graph-Based Approach*, in PROC. FOURTH WORKSHOP ON ONLINE ABUSE & HARMS 191 (2020).

31. *Supra* note 2.

D. The Disclosure Penalty

The final layer supplies motive. The crowd holding the button is not indifferent to the question it is asked; a substantial and replicated literature finds active penalty. Workers who use AI are rated lower on competence and motivation, the penalty carries into hiring evaluations, and users correctly anticipate it.[32] Across sixteen preregistered experiments and 27,491 participants, evaluations of writing dropped when participants believed a model wrote it or helped write it, the effect was mediated by perceived authenticity, and it survived every mitigation the authors drew from prior work.[33] Disclosing AI use erodes trust across tasks and roles, whether the disclosure is voluntary or mandated.[34] The label alone does damage: marking a headline as AI-generated lowered its perceived accuracy and readers' willingness to share it even where the headline was true and human-written.[35] The nuance in this literature cuts no slack for the instrument: where researchers decompose the effect, much of it is favoritism toward known human authorship rather than pure aversion,[36] which means the penalty attaches at disclosure, and attaches to assisted work as such.

The moderator evidence is reported here rather than waved away, because it converts. The trust penalty attenuates, without disappearing, among evaluators with favorable technology attitudes;[37] higher AI literacy softens and occasionally reverses the perception loss;[38] disclosure framed as human control over the tool, rather than as machine share, closes the gap in consumer

---

32. Jessica A. Reif, Richard P. Larrick & Jack B. Soll, *Evidence of a Social Evaluation Penalty for Using AI*, 122 Proc. Nat'l Acad. Scis. e2426766122 (2025).

33. Manav Raj et al., *The Artificial Intelligence Disclosure Penalty: Humans Persistently Devalue AI-Generated Creative Writing*, 155 J. Experimental Psych.: Gen. 896 (2026).

34. Oliver Schilke & Martin Reimann, *The Transparency Dilemma: How AI Disclosure Erodes Trust*, Org. Behav. & Hum. Decision Processes, art. 104405 (2025).

35. Sacha Altay & Fabrizio Gilardi, *People Are Skeptical of Headlines Labeled as AI-Generated, Even If True or Human-Made, Because They Assume Full AI Automation*, 3 PNAS Nexus pgae403 (2024).

36. Yunhao Zhang & Renée Gosline, *Human Favoritism, Not AI Aversion*, 18 Judgment & Decision Making e41 (2023).

37. Schilke et al., *supra* note 34 (within-paper meta-analysis).

38. Hiroki Nakano et al., *Understanding Reader Perception Shifts upon Disclosure of AI Authorship*, in PROC. 31ST INT'L CONF. ON INTELLIGENT USER INTERFACES (2025).

content;[39] and source transparency counteracts the label penalty in news.[40] Every one of those moderators operates through a disclosure surface and an informed evaluator. The button supplies neither. The flag fires on register before any authorial statement is seen, the reporting member is never shown the writer's disclosure, and the platform's design gives the disclosure no surface at all. The moderation literature is therefore the efficacy case for the remedy Part V specifies, published in advance by the researchers who mapped the penalty's limits. And the temporal direction runs against the platform. Community-level restriction of AI content is accelerating: the number of communities adopting rules against AI content on a major discussion platform more than doubled in a single year, on stated grounds of quality and authenticity,[41] and the strongest recent experimental lines, sixteen preregistered experiments published in 2026 and thirteen in 2025, find the penalty persisting against the very interventions drawn from the earlier, more optimistic literature.[42] A January 2026 laboratory study adds a design detail the remedy already anticipates: the trust penalty tracked the verbosity of the public label, with detailed disclosures depressing trust and subscriptions where one-line disclosures did not, which is why the channel Part V specifies discloses to the platform and leaves any public display to the member's election.[43] The writer whose accessibility disclosure names the assistance has prepaid the penalty the button now collects.

E. The Stack

Assemble the layers. The platform assigned a discrimination task that humans measurably cannot perform, to raters whose judgments skew against the prose registers of protected classes,

---

39. Martin Haupt, Jan Freidank & Alexander Haas, *Consumer Responses to Human-AI Collaboration at Organizational Frontlines: Strategies to Escape Algorithm Aversion in Content Creation*, 19 Rev. Managerial Sci. 377 (2024).

40. Benjamin Toff & Felix M. Simon, *"Or They Could Just Not Use It?": The Dilemma of AI Disclosure for Audience Trust in News*, 30 Int'l J. Press/Politics 881 (2024).

41. Travis Lloyd et al., *AI Rules? Characterizing Reddit Community Policies Towards AI-Generated Content*, in Proc. 2025 CHI Conf. on Hum. Factors in Computing Sys. (2025).

42. Raj et al., *supra* note 33; Schilke & Reimann, *supra* note 34.

43. Pooja Prajod, Hannes Cools & Thomas Röggla, *Full Disclosure, Less Trust? How the Level of Detail About AI Use in News Writing Affects Readers' Trust*, arXiv:2601.09620 (2026) (N = 40).

who carry a documented penalty reflex against disclosed assistance, through an instrument that records only accusations and admits no ground truth, and it trains the classifier that prices distribution on the result. Each layer is independently established in the record above. The layers compound. The doctrinal question the remaining Parts answer is whether the law of disability discrimination reaches a design like this. It does, and the remedy it supports is smaller than the platform's likely fears and larger than its current obligations.

IV. The Disability Frame

A. Impact Without Intent

Nothing in the analysis requires the platform to have aimed at disabled writers, and the Article does not allege that it did. The governing doctrine has never required aim. *Alexander v. Choate* holds that Section 504 reaches at least some conduct that discriminates in effect, and it fixes the touchstone: the protected class must be afforded "meaningful access" to the benefit the covered entity offers, and the benefit may not be defined in a way that effectively denies that access.[44] The Ninth Circuit's decision in *Payan* confirms that private plaintiffs may enforce that impact principle: *Alexander v. Sandoval* does not foreclose private disparate impact claims under Title II and Section 504, and the court identified the theories under which a facially neutral practice that burdens the class is actionable.[45] The circuits divide on that enforcement question, and this Article does not pretend otherwise; the Sixth has read *Choate*'s reservation the other way. The division matters less here than it first appears, because the Title III hooks pleaded below are statutory effects text, methods of administration and screen-out, and do not depend on the implied-right question the Ninth resolved. The 2026 sequel trims damages, and it trims them in a way that matters here: emotional distress is out under the extension of *Cummings*, and compensatory damages for lost opportunities, the economic value of professional distribution

44. *Alexander v. Choate*, 469 U.S. 287, 299–301 (1985).

45. *Payan v. L.A. Cmty. Coll. Dist.*, 11 F.4th 729, 738–40 (9th Cir. 2021); *see Alexander v. Sandoval*, 532 U.S. 275 (2001); *but see Doe v. BlueCross BlueShield of Tenn., Inc.*, 926 F.3d 235, 241 (6th Cir. 2019) (concluding that § 504 does not prohibit disparate-impact discrimination).

denied, expressly remain.[46] The professional platform is where distribution is the benefit; the damages theory that survives is the one this injury produces.

Title III writes the impact principle into text three times over. A public accommodation may not, directly or through arrangements, "utilize standards or criteria or methods of administration that have the effect of discriminating on the basis of disability."[47] It may not impose eligibility criteria that "screen out or tend to screen out" individuals with disabilities from full and equal enjoyment of its services.[48] And it must make reasonable modifications to policies, practices, and procedures where necessary to afford its services to individuals with disabilities.[49] The flag pipeline is a method of administration: a standing procedure by which the platform administers the distribution of member speech. The register criterion the crowd applies, and the classifier will automate, is a criterion that tends to screen out. The statutory hooks fit the design without strain.

B. The Class, the Mitigating Measure, and the Penalized Trace

The class is writers whose disabilities are mitigated through assistive technology that touches text. The statute defines disability functionally, and writing, speaking, and communicating are enumerated major life activities.[50] The ADA Amendments Act then adds the provision that gives this Article its sharpest interpretive frame: in determining whether an individual is disabled, the ameliorative effects of mitigating measures are disregarded, and the statute's illustrative list of mitigating measures begins with assistive technology.[51] The provision is definitional; it prescribes how disability is assessed and prohibits nothing by itself, and this Article deploys it as the statute's stated policy while the operative prohibitions remain § 12182(b)(1)(D) and (b)(2)(A). As policy, it is pointed: Congress ordered decisionmakers to look past the assistive

46. *Payan v. L.A. Cmty. Coll. Dist.*, 169 F.4th 971, 977–79 (9th Cir. 2026); *Cummings v. Premier Rehab Keller, P.L.L.C.*, 596 U.S. 212 (2022).

47. 42 U.S.C. § 12182(b)(1)(D).

48. 42 U.S.C. § 12182(b)(2)(A)(i).

49. 42 U.S.C. § 12182(b)(2)(A)(ii).

50. 42 U.S.C. § 12102(1), (2)(A).

51. 42 U.S.C. § 12102(4)(E)(i).

technology to protect the person. The platform's instrument does the reverse: it detects the trace of the mitigating measure and penalizes the person for it. The dictation pipeline exists because sustained typing is what the impairment restricts; the model-touched surface of the text is the visible residue of the accommodation. An instrument that prices distribution on that residue prices distribution on disability, one step removed, and the step is the accommodation itself.

Causation runs through Part III without a gap. The register features the crowd treats as machine tells, disciplined transitions, formal connectives, consistent mechanics, are the features assistive pipelines produce and the features the measured crowd penalizes in quality writing generally.[52] The disclosure penalty literature closes the motivational loop: the population holding the button punishes disclosed assistance as such.[53] A plaintiff class does not need the platform's logs to plead plausibility; the public record supplies the mechanism at every link.

C. Within-Class Impact and the Comparator

Choate cautions that not every disparate effect states a claim; the inquiry is whether the class is denied meaningful access to the benefit actually offered.[54] The benefit the platform offers every member is the same: distribution of professional speech to an audience assembled by the platform, on terms administered by the platform. The comparator is the nondisabled member whose unassisted prose carries no machine trace and therefore never enters the accusation pool on the basis of register. The disabled member's speech enters that pool because of the accommodation, is demoted on accusation without notice, and can exit the demotion only through the paid channel Part VI addresses. That is a denial of meaningful access to the offered benefit, and Part II.C supplies what impact doctrine treats as the finishing argument: an equally effective, less discriminatory alternative was available, was already built, and is still running. The platform need not tolerate content its members dislike; the "not interested" control lets

52. *Supra* notes 17, 21 and accompanying text.
53. *Supra* notes 32–33 and accompanying text.
54. *Choate*, 469 U.S. at 299–306.

every member curate a feed. What the platform added was the accusatory instrument, and the increment between the two is where the statutory violation lives.

V. Title III Coverage and the Procedural Remedy

A. The Platform as Public Accommodation

Title III's coverage question, whether a web platform is a place of public accommodation, carries a familiar split. The First Circuit held in *Carparts* that public accommodations are not limited to physical structures, reasoning that a statute enumerating "travel service" and "other service establishment" would be absurd if it protected customers who walked in the door and abandoned those who called or wrote.[55] The Seventh Circuit read the statute the same way.[56] Other circuits require a nexus to a physical place,[57] but the Ninth Circuit's own application shows how little the nexus demands where the online service is the channel through which a covered entity's services flow: *Robles* held the ADA applies to the website and app of a pizza chain because they connect customers to the goods and services of the covered places.[58]

The Article states both routes and needs only one. Under the service-establishment reading, a professional networking platform that sells memberships, advertising, recruiting services, and distribution is an "other service establishment" without more.[59] Under the nexus reading, the correct defendant is the parent, and the stores are why. Microsoft Corporation operates physical places of public accommodation; its flagship retail floor is a sales establishment in the classic sense.[60] Title III then forbids that covered entity to discriminate in the full and equal enjoyment of its services "directly, or through contractual, licensing, or other

55. *Carparts Distrib. Ctr., Inc. v. Auto. Wholesaler's Ass'n of New Eng., Inc.*, 37 F.3d 12, 19–20 (1st Cir. 1994).

56. *Doe v. Mut. of Omaha Ins. Co.*, 179 F.3d 557, 559 (7th Cir. 1999).

57. *See, e.g.*, *Weyer v. Twentieth Century Fox Film Corp.*, 198 F.3d 1104, 1114–15 (9th Cir. 2000); *Ford v. Schering-Plough Corp.*, 145 F.3d 601, 612–14 (3d Cir. 1998); *Parker v. Metro. Life Ins. Co.*, 121 F.3d 1006, 1010–11 (6th Cir. 1997) (en banc).

58. *Robles v. Domino's Pizza, LLC*, 913 F.3d 898, 905 (9th Cir. 2019).

59. 42 U.S.C. § 12181(7)(F).

60. 42 U.S.C. § 12181(7)(E); *see, e.g.*, Microsoft Experience Center, New York, N.Y. (documentation on file with author).

arrangements,"[61] and the professional platform is a wholly owned line of business the parent markets, bundles, and sells alongside its covered operations. The claim in that configuration does not ask a court to hold the subsidiary covered by affiliation, a step no case has taken; it holds the covered parent to the statute for what the parent does through its arrangement. The application is novel and the entity-separateness objection is real, which is why the service-establishment route carries the Part on its own where the nexus route fails. The coverage question deserves candor, and candor permits confidence: the claim survives under either line, and the split itself is the subject of a developed literature this Article does not need to resolve.

B. The Modification, Specified

The remedy is a reasonable modification of policies, practices, and procedures,[62] and precision here is the argument. Three procedures:

First, a disclosure channel: a member may register, once, that their writing is produced through assistive technology, with no obligation to specify diagnosis and no public display absent the member's election. Second, notice: a member whose post's distribution is reduced on authenticity flags is told so, in the analytics surface the platform has already announced it is testing.[63] Third, review: on the disclosed member's appeal, a human decides whether the penalty stands, and flags on disclosed accounts do not enter classifier training while review is pending.

The individualized character of the inquiry is the doctrine working as designed: *PGA Tour v. Martin* requires the covered entity to consider the particular individual's circumstances rather than administer the rule in gross.[64] The fundamental-alteration defense has no purchase. The platform operates appeal flows for other report categories, operates preference controls, and has announced the notice surface itself; the modification asks it to connect components it has already built. Part X adds the finishing fact: the platform's European operations are already legally

61. 42 U.S.C. § 12182(b)(1)(A)(i).

62. 42 U.S.C. § 12182(b)(2)(A)(ii).

63. *Supra* note 2; *supra* Part II.A.

64. *PGA Tour, Inc. v. Martin*, 532 U.S. 661, 688 (2001).

obligated to provide the statement of reasons and human review this modification describes, so the burden argument reduces to a refusal to run, in one market, the process it runs in another.

C. What the Modification Does Not Do

The modification alters no member's content. It removes no member's speech. It compels no member's speech. It does not forbid any member to press any button, and it does not require the platform to referee whether any post is in fact machine-made. Its entire operation is addressed to the platform's own penalty procedure: who gets told, who gets a hearing, and what the platform's own classifier may ingest while a hearing is pending. Every word of that design is deliberate, and Part VIII collects the doctrinal payment.

VI. The Surcharge

The regulation is one sentence and it decides the Part: a public accommodation "may not impose a surcharge on a particular individual with a disability or any group of individuals with disabilities to cover the costs of measures" required to provide nondiscriminatory treatment.[65] The baseline of nondiscriminatory treatment, Parts IV and V establish, is distribution unreduced by register accusations that proxy the accommodation. The platform's paid promotion products restore exactly the quantity the flag pipeline takes: money in, distribution back. For the nondisabled member, promotion buys amplification above baseline. For the flagged disabled member, the same purchase buys the baseline back. The identical price schedule conceals a discriminatory transaction: the class pays to reach the position others occupy for free, which is the surcharge in its regulatory definition, the cost of the equalizing measure shifted onto the individual it protects.

Two of the platform's own facts close the escape routes. The company sells promoted distribution of professional thought-leadership content as a product line, which forecloses any argument that paid restoration of an expert post's reach fundamentally alters the service; the

65. 28 C.F.R. § 36.301(c); *accord* 28 C.F.R. § 35.130(f) (public entities).

service is offered for sale to everyone.[66] And the flag's reach penalty operates automatically and invisibly while the paid channel is marketed continuously, a structure in which the platform profits from the very demotion the class must pay it to undo. The Article does not need to attribute that structure to design. The regulation does not ask why the charge exists; it asks who pays it, and for what. One candor note completes the Part: no court has yet applied the surcharge rule to paid restoration of algorithmically reduced distribution, and the claim is derivative by design, operative once Parts IV and V establish the baseline it prices.

VII. Section 504 and the Corporate Parent

A. Assistance, Not Procurement

Section 504 reaches any program or activity receiving federal financial assistance.[67] The threshold discipline is the line between assistance and procurement. The Federal Grant and Cooperative Agreement Act separates the instruments by purpose: contracts acquire goods and services for the government's direct benefit; grants and cooperative agreements transfer value to carry out a public purpose.[68] The case law enforces the line: payment under a procurement contract is compensation, and compensation is not federal financial assistance.[69] The platform's parent sells the government software and cloud services at scale; none of that matters here. What matters is the narrower and better fact.

B. The Parent as Recipient

The parent holds federal awards classified as assistance: a completed Department of Defense award of approximately $6.5 million associated with its quantum computing research, an active Department of Agriculture research award of approximately $1 million, and active Department of Defense cooperative agreements with the National Center for Manufacturing Sciences

66. Thought Leader Ads solicitation received by the author (on file with author).
67. 29 U.S.C. § 794(a).
68. 31 U.S.C. §§ 6303–6305.
69. *Jacobson v. Delta Airlines, Inc.*, 742 F.2d 1202, 1209 (9th Cir. 1984).

consortium through which sub-awards flow, all reflected in the public award record.[70] Recipient status is the statutory trigger, and the Supreme Court's cases mark its boundaries from both sides: entities that merely benefit from assistance extended to others are not covered,[71] while entities to which assistance is extended are, and the inquiry follows the assistance rather than the org chart's labels.[72] The parent is a named recipient, not a downstream beneficiary. The temporal posture is stated plainly rather than papered over: present-tense receipt runs through the live Department of Agriculture award and the active consortium instruments, while the completed Department of Defense award supplies the history of direct assistance rather than a current hook.

C. Scope, and the Question Stated Openly

Coverage scope is where candor is owed. For a covered corporation, "program or activity" extends to all of the operations of the entity where the assistance is extended to the corporation as a whole, and is otherwise confined to the assisted operation or facility.[73] The parent's directly operated AI products sit inside the parent's own operations, and the argument for whole-entity coverage rises or falls on the structure of the awards. The platform, by contrast, is a separately incorporated subsidiary. The dominant reading stops recipient status at the corporate boundary absent assistance to the subsidiary itself; a broader reading, grounded in the restoration statute's history, would reach wholly owned subsidiaries operated as integrated lines of business. This Article states the question as open because it is open, and structures the claim accordingly: Section 504 is pleaded against the parent for the parent-operated surfaces on which the same

---

70. USAspending.gov award records (screen captures on file with author): U.S. Dep't of Agric., other financial assistance award, FAIN 5830222032, Assistance Listing 10.001 (Agricultural Research Basic and Applied Research) (recipient Microsoft Corporation, Redmond, Wash.) (in progress); U.S. Dep't of Def., other financial assistance award, FAIN H982300830001, Assistance Listing 12.901 (Mathematical Sciences Grants) (recipient Microsoft Corporation) (completed); U.S. Dep't of Def., cooperative agreements, FAINs HQ00341520007 and HQ00342020007, Assistance Listing 12.225 (Commercial Technologies for Maintenance Activities Program) (recipient National Center for Manufacturing Sciences, Inc., the consortium intermediary) (in progress); Microsoft Corporation recipient profile, UEI INLWM5KDS7R9, `https://www.usaspending.gov/recipient/dd77b7c3-663e-cb91-229f-5766a50e9b7f-P/all` (last visited Aug. 13, 2026).

71. *U.S. Dep't of Transp. v. Paralyzed Veterans of Am.*, 477 U.S. 597, 605–07 (1986).

72. *NCAA v. Smith*, 525 U.S. 459, 468 (1999).

73. 29 U.S.C. § 794(b)(3)(A)–(B).

detection-and-penalty logic runs, while the Title III analysis of Parts IV through VI, which needs no federal money, carries the platform itself. Two straws, independently sufficient burdens of proof, one class.

VIII. Section 230 and the First Amendment

A. What the Claim Targets

Section 230(c)(1) forbids treating a provider as "the publisher or speaker of any information provided by another information content provider."[74] The statute's protection, moreover, is a defense to liability rather than an immunity from suit. The Supreme Court drew that line in 2026, and the Ninth Circuit applied it to Section 230 within days of this writing, dismissing interlocutory appeals from the denial of Section 230 motions because a defense to liability is fully reviewable after final judgment.[75] A platform that invokes the statute litigates it.

The design challenged here contains third-party information in exactly one place: the flags. The claim does not seek to hold the platform liable for what any flagger expressed. It targets the platform's own instrument, the platform's own classifier, and the platform's own administration of distribution, and it seeks a remedy addressed solely to the platform's own procedure. Where immunity attaches and where it does not now varies by circuit, and the Article maps the terrain as it stands in August 2026 rather than as any party would prefer it.

B. Three Circuits, Three Configurations

The Third Circuit holds that a platform's curation algorithm is the platform's own expressive product after *Moody*, and that Section 230 therefore does not reach claims aimed at the curation

74. 47 U.S.C. § 230(c)(1).

75. *Pers. Injury Plaintiffs v. TikTok, LLC*, No. 24-7312 (9th Cir. Aug. 10, 2026) (holding that Section 230 provides a defense to liability rather than immunity from suit) (applying *GEO Grp., Inc. v. Menocal*, 607 U.S. 438, 445, 449 (2026)).

itself.[76] In that configuration the demotion pipeline is first-party conduct and the analysis proceeds directly to the merits. The platform's likely rejoinder, that a crowd vote makes it a passive receptacle, fails on its own facts: the platform wrote the button, weighted the votes, priced the penalty, and trains the classifier.

The Ninth Circuit, at panel level, runs the other way: matching users with content is publishing conduct within the statute even when unrequested,[77] over a concurrence that would treat the recommendation algorithm as the platform's own distinctive expressive product.[78] In that configuration the claim survives through structure, Part VIII.C.

The Fifth Circuit, three weeks before this writing, rejected the premise that the two shields trade off at all. Confronting the argument that *Moody*'s recognition of first-party editorial interests takes curation outside Section 230, the court declined to read *Moody* as rendering the statute avoidable at all, reasoning that speech can be protected by the First Amendment and simultaneously covered by the statute's bar on treating the speaker as a publisher.[79] The shields stack, and the operative test in that configuration is functional: whether the asserted duty "would necessarily require an internet company to monitor, alter, or remove third-party content."[80] A dissent in the same case carries the first-party view into the Fifth Circuit's records.[81] The Supreme Court, for its part, has decided nothing: its one encounter ended in a per curiam vacatur that left the statute's scope untouched.[82]

Now apply the strictest of the three tests to the Part V modification. A disclosure channel requires the platform to accept a registration. Notice requires it to surface its own penalty decision to the penalized member. Review requires one of its employees to decide an appeal of

---

76. *Anderson v. TikTok, Inc.*, 116 F.4th 180, 183–84 (3d Cir. 2024) (TikTok's algorithm, curating and recommending, "was TikTok's own expressive activity, and thus its first-party speech"); *see Moody v. NetChoice, LLC*, 603 U.S. 707, 738 (2024).

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

78. *Id.* (R. Nelson, J., concurring).

79. *Computer & Commc'ns Indus. Ass'n v. Paxton*, No. 24-50721 (5th Cir. July 24, 2026).

80. *A.B. v. Salesforce, Inc.*, 123 F.4th 788, 794–95 (5th Cir. 2024) (quoting *Force v. Facebook, Inc.*, 934 F.3d 53, 83 (2d Cir. 2019) (Katzmann, C.J., concurring in part and dissenting in part)).

81. *Paxton*, *supra* (Ho, J., concurring in part and dissenting in part).

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

its own procedure, and the training carve-out governs what the platform's own classifier ingests. None of the three, and no combination of them, requires the platform to monitor, alter, or remove any third party's content. The duty survives the Fifth Circuit's test with the immunity fully intact, survives the Ninth's on the same structure, and in the Third the immunity never attaches to the curation conduct at all. The remedy was designed for the worst configuration and therefore holds in every configuration.

C. No Exemption, and None Needed

The claim structure matters because the exemption does not exist. A district court squarely rejected the argument that the ADA's status as a civil rights statute carves it out of Section 230,[83] and this Article does not renew it. The surviving route is the one the Ninth Circuit's own line marks: immunity fails where the asserted duty can be satisfied without any change to third-party content,[84] which is the Part V design restated as a holding. The sentence the Article defends is exact: there is no disability exemption in Section 230, and the claim survives anyway, because it targets the platform's own design and asks only for process.

D. Weaponized Flagging

One further scenario requires its own doctrine: coordinated flagging campaigns aimed at a disclosed writer by parties hostile to the writer's work. The flags in that scenario are unmistakably third-party expression, and no claim here treats the flaggers' expression as the wrong. The instructive analogy is the housing rule of *Wetzel*: a covered entity with actual notice of third-party harassment of a protected resident, holding an arsenal of instruments to address it, that chooses deliberate indifference, is directly liable for its own inaction, and the third parties'

---

83. *Nat'l Ass'n of the Deaf v. Harvard Univ.*, 377 F. Supp. 3d 49 (D. Mass. 2019) (rejecting the argument that the ADA and Rehabilitation Act displace § 230); *see also Sikhs for Justice, Inc. v. Facebook, Inc.*, 144 F. Supp. 3d 1088 (N.D. Cal. 2015) (civil rights claim treated no differently under § 230), *aff'd*, 697 F. App'x 526 (9th Cir. 2017).

84. *HomeAway.com, Inc. v. City of Santa Monica*, 918 F.3d 676, 682–83 (9th Cir. 2019) (no immunity where the duty "does not require" monitoring or editing third-party content); *In re Apple Inc. App Store Simulated Casino-Style Games Litig.*, 625 F. Supp. 3d 971 (N.D. Cal. 2022) (duties independent of the content of third-party listings survive).

authorship of the harassment is no defense.[85] The analogy is offered with its limits stated: *Wetzel* is a Fair Housing Act case about a custodial residential setting, and this Article claims only its structure, notice, control, and indifference as the platform's own conduct. The disclosure channel of Part V converts the structure into an operating standard: after disclosure, the platform is on notice; the flag pipeline is the platform's own control instrument; what it does next is its own act.

E. The First Amendment, Directly

Moody establishes that a platform's expressive curation choices receive First Amendment protection.[86] The modification does not touch them. It compels no message, prescribes no ranking, and forbids no editorial judgment; the platform may demote whatever its editorial voice demotes, subject only to telling a disclosed disabled member that it happened and hearing an appeal of its own procedure. Process obligations attached to an internal penalty pipeline are conduct regulation of a kind courts have sustained against expressive enterprises, newspapers included, where the obligation leaves the editorial product itself untouched. The Article meets the stronger objection honestly: if a court someday holds that the act of demoting register-flagged prose is itself protected editorial expression immune from process, that holding would prove the Article's descriptive thesis in full, that the platform claims an expressive identity built on penalizing the accommodation, and the legislative recommendations of Part X become the operative agenda.

IX. The Deception Lane

A. The Promise Doctrine

One family of claims never meets Section 230 at all. *Barnes* holds that promissory liability arises from the promise, and that enforcing a specific undertaking does not treat the promisor

85. *Wetzel v. Glen St. Andrew Living Cmty., LLC*, 901 F.3d 856, 861–65 (7th Cir. 2018) (direct liability where the operator had actual notice of tenant-on-tenant harassment, held an arsenal of remedial incentives and sanctions under its own agreements, and chose inaction).

86. *Moody*, 603 U.S. at 738.

as a publisher.[87] The Ninth Circuit's 2025 refinement fixes the boundary with a drafting instruction embedded in it: a general representation that a platform maintains a "safe and secure environment" is a description of moderation policy and is immunized; a specific promise, take down these profiles, unmask these senders, is enforceable.[88] The count that follows is drafted to the specific side of that line.

B. The Paid-Distribution Transaction

The platform sells distribution. A member pays to promote an identified post to a quantified audience; the seller's own systems, meanwhile, demote the class of posts the member writes, on accusations the seller solicits, through the pipeline of Part II. Two theories attach. Under Section 5 of the FTC Act, the sale of amplification to a writer whose organic distribution the seller is simultaneously and undisclosedly suppressing on register is a deceptive practice: the transaction's premise, that payment buys position above an honest baseline, is false as to the flagged class.[89] Under the Illinois Consumer Fraud and Deceptive Business Practices Act, the same facts state a claim for deception in trade with intent that the buyer rely, and the state statute supplies the private right the federal one lacks.[90] The state count presupposes its forum facts, an Illinois purchaser and a transaction reaching Illinois commerce, pleaded in the ordinary course. The knowledge element is not a pleading problem; it is the platform's own press. The company publicly quantifies the automated inauthentic activity it blocks at massive scale, which is a public representation that it knows, measures, and controls the integrity of the distribution it sells.[91]

The design is testable, and the Article commits to the test rather than the insinuation: promoted posts' engagement profiles can be compared against organic baselines, reactor

87. *Barnes v. Yahoo!, Inc.*, 570 F.3d 1096, 1107–09 (9th Cir. 2009).

88. *Doe v. Grindr Inc.*, 128 F.4th 1148 (9th Cir. 2025) (holding that a general statement that an app is designed to create a safe and secure environment describes moderation policy rather than making a specific promise); *Est. of Bride v. Yolo Techs., Inc.*, 112 F.4th 1168, 1173, 1178–79 (9th Cir. 2024).

89. 15 U.S.C. § 45(a).

90. 815 ILCS 505/2.

91. Srinivasan, quoted in Fortune, *supra* note 2 (the company "detects and blocks hundreds of thousands of automated slop comment attempts every day" and has "prevented 'billions of other automation attempts (posting at scale, slop) in the last couple months alone"').

account patterns coded, and the integrity of the sold product measured from public surfaces. The empirical companion to this Article specifies the protocol; the doctrinal point here is narrower and complete: whatever that audit shows, the promise framework of *Barnes* and *Grindr* governs it, and none of it passes through Section 230.

X. The Digital Services Act Counterpoint

The European Union has already legislated on the exact conduct. Three provisions measure the shipped design.

Article 17 requires a provider to give an affected recipient a clear and specific statement of reasons for "any restrictions of the visibility" of the recipient's content, demotion included, identifying the facts, the automated means used, and the redress available.[92] The shipped design provides the affected author nothing; the notice surface exists only as an announced test.[93] Article 20 requires an internal complaint-handling system for those decisions, and forbids deciding complaints "solely on the basis of automated means": a human must be in the loop the modification of Part V describes.[94] Article 24(5) requires the statements of reasons to be filed to the Commission's public database, which converts compliance into an auditable public record.[95]

The comparative point is not that European law governs an American claim. It is that the platform's own European operations are already legally required to deliver notice, reasons, and human review for visibility restrictions, which extinguishes any argument that the Part V modification is infeasible, fundamentally altering, or unduly burdensome. The platform has built the process. The question this Article poses is only whether American disability law entitles the American disabled writer to the process the platform already runs for everyone in Europe. On the doctrine assembled above, it does; and where any court concludes otherwise, the DSA supplies the drafted, operating, auditable template for the legislative correction.

---

92. Regulation (EU) 2022/2065, art. 17, 2022 O.J. (L 277) 1 (Digital Services Act).
93. *Supra* Part II.A.
94. *Id.* art. 20.
95. *Id.* art. 24(5).

XI. Conclusion

A platform built an instrument that collects accusations of inauthenticity, only accusations, from a crowd the record shows to be inaccurate, demographically skewed, and primed to punish disclosed assistance, and it wired the accusations to distribution and to the training of the system that will automate them. The writers who pay first are the ones whose access to written professional speech runs through the assistance being accused. The law does not require the platform to referee authenticity. It requires process before penalty, and it forbids selling the exit. Congress ordered the law to look past the mitigating measure to see the person; the platform built a machine that looks past the person to punish the mitigating measure. The platform built the adequate alternative years ago and still operates it, and its European surfaces already deliver the notice, the reasons, and the human review that its American disabled members are told would break the product. The distance between what it built and what it shipped on July 30 is the measure of the violation.

References

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

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.

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

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

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.

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

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

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

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

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.

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

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.

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

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

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

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).

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

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

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

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

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).

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.

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

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

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).

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*.

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

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*.

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

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

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.

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).

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.

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

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

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

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.

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

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.

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

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.

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

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.

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).

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.

❧ ❧