The Distributor's Burden:

Platform Verification Duties, Privacy-Preserving Age Assurance, and the Constitutional Ceiling on Online Child-Safety Design

Travis Gilly*

Working draft, July 2026 (v2)

Abstract

The public argument over children and online platforms is being fought on the wrong axis. When a platform frames a proposed media-prominence rule as a threat to open recommendation, or a legislature answers a stream of harmful content by demanding that every user prove an age, both moves aim the burden of child protection at the wrong party. The classification that actually decides whether a video reaches a child runs, on the largest platforms, on creator self-report, and self-report is a sieve. A May 2026 field study found that roughly a third of surveyed children had already bypassed age checks, some by drawing a moustache with a makeup pencil and some by pointing a camera at an animated video-game character.

This paper argues that the burden belongs to the distributor. It first clears two wrong answers. User-side mass age verification collides with a constitutional ceiling: *Free Speech Coalition v. Paxton* sustained age checks only for material obscene as to minors, and the district courts have spent the year since enjoining every attempt to stretch that authorization across general-audience platforms. Government content-prominence mandates fare no better, because *Moody v. NetChoice* holds that curation is protected editorial judgment and that a State may not conscript a platform to rebalance the speech market. The paper then identifies the doctrinal engine a distributor duty can run on: the symmetry between *Moody* and *Anderson v. TikTok*. If the curated feed is the platform's own expressive product when the platform claims constitutional protection, it is the platform's own first-party conduct when the question is accountability, and Section 230 does not stand in the way. The paper's contribution is a distributor verification duty, scaled to platform capacity, that requires a platform to verify its own child-directed classifications rather than trust a creator's checkbox, paired with a narrow adult tier gated by privacy-preserving age assurance rather than surveillance. The whole-video scanning that platforms already run for advertiser safety proves the capability

*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.

exists; the drafting lessons of *NetChoice v. Bonta* and the vagueness cases show how to write the duty so it regulates factual classification rather than compelled opinion. The design threads the constitutional ceiling under *Paxton*, honors editorial autonomy under *Moody*, rests accountability on *Anderson*, keeps the collect-nothing property of a zero-knowledge age token, and confines the identification residual to the adult tier alone.

Keywords: child safety, age assurance, zero-knowledge proof, *Free Speech Coalition v. Paxton*, *Moody v. NetChoice*, *Anderson v. TikTok*, Section 230, duty of care, content moderation, First Amendment, platform design.

Contents

I. Introduction: The Wrong Axis

A platform and a legislature can disagree about children online while sharing a mistake. When YouTube urged its creators to oppose a British proposal that would require video-sharing services to give prominence to public-service broadcasters, it cast the child-protection question as a fight over which channels an algorithm elevates.[1] When a legislature answers a stream of material harmful to children by requiring every visitor to prove an age at the door, it aims the same burden at the audience. Both moves point away from the party that decides what reaches a child, and that party is the platform.

The decision that matters is a classification. On the largest video service, whether a given upload is treated as directed to children turns in the first instance on the uploader's own declaration, a self-report artifact of federal children's-privacy law.[2] A self-report gate has a predictable failure mode: the actor with the most reason to misdeclare is the one the gate is meant to stop, and nothing forces a second look. A field study published in May 2026 measured the audience-side counterpart of the same failure. Surveying more than a thousand children, the nonprofit Internet Matters reported that roughly a third had bypassed age checks within a two-month window, with methods ranging from a false birthdate to drawing facial hair with a makeup pencil to pointing a camera at an animated video-game character so that an age-estimation model read an adult.[3]

---

1. The proposal appears in the United Kingdom's green paper on the future of public-service media. DEP'T FOR CULTURE, MEDIA & SPORT, *Watch This Space: A New Strategic Direction for UK Media* (2026). For the platform's public response, see James O'Toole, *YouTube Urges Creators To Fight Proposed UK Algorithm Changes*, DEXERTO (July 4, 2026), `https://www.dexerto.com/entertainment/youtube-urges-creators-to-fight-proposed-uk-algorithm-changes-3383111/` (last visited July 5, 2026).

2. The "made for kids" designation implements the Children's Online Privacy Protection Act, 15 U.S.C. §§6501–6506, and its implementing rule, 16 C.F.R. pt. 312. The designation became central to the platform's operations through the 2019 enforcement settlement resolving alleged COPPA violations on YouTube. See Press Release, Fed. Trade Comm'n, *Google and YouTube Will Pay Record $170 Million for Alleged Violations of Children's Privacy Law* (Sept. 4, 2019), `https://www.ftc.gov/news-events/news/press-releases/2019/09/google-youtube-will-pay-record-170-million-alleged-violations-childrens-privacy-law` (last visited July 5, 2026).

3. INTERNET MATTERS, *The Online Safety Act: Are Children Safer Online?* (2026). For contemporaneous reporting of the study's findings, see Zack Whittaker, *Some Kids Are Bypassing Age-Verification Checks With a Fake Mustache*, TECHCRUNCH (May 6, 2026), `https://techcrunch.com/2026/05/06/some-kids-are-bypassing-age-verification-checks-with-a-fake-mustache/` (last visited July 5, 2026).

This paper argues that the burden belongs to the distributor and shows what a lawful version of that allocation looks like. Part II clears the two wrong answers, mass user-side verification and government prominence mandates, and locates the constitutional ceiling that constrains any child-safety design. Part III identifies the doctrinal engine of a distributor duty: the symmetry between the feed as protected expression and the feed as accountable first-party conduct. Part IV sets out the proposal, a distributor verification duty scaled to capacity and drafted within the compelled-speech and vagueness constraints the courts have already marked, paired with privacy-preserving age assurance for a narrow adult tier. Part V shows how the design satisfies the constitutional, privacy, access, and competition constraints at once. Part VI marks the limits and the open questions.

II. Two Wrong Answers

A. Mass User-Side Verification and the Constitutional Ceiling

1. The Paxton authorization and its limits

The reflexive answer to harmful content is to verify everyone. The Supreme Court's most recent word on age verification authorizes far less. In *Free Speech Coalition, Inc. v. Paxton*, the Court sustained a Texas statute requiring age verification on commercial sites a substantial share of whose content is sexual material harmful to minors, holding that such a requirement draws intermediate scrutiny because it only incidentally burdens the protected speech of adults.[4] The Court reasoned that no person, adult or child, has a First Amendment right to reach material that is obscene as to minors without first submitting proof of age, so the incidental character of the burden on adults follows.[5] That reasoning displaced, for this category, the strict-scrutiny approach of the early internet cases, which had struck down broad federal attempts to wall off indecent online speech

4. Free Speech Coal., Inc. v. Paxton, 606 U.S. 461 (2025).

5. *Id.* The Court grounded the State's authority in the variable-obscenity doctrine of Ginsberg v. New York, 390 U.S. 629, 638 (1968), under which a State may adjust the definition of obscenity to the perspective of minors and restrict, on rational-basis review, material that appeals to the prurient interest of minors, depicts sexual conduct in a way patently offensive for minors, and lacks serious value for minors.

and had insisted that the government may not reduce the adult population to reading only what is fit for children.[6]

The reach of the holding is the point that matters here, and the opinion is careful about it. *Paxton* addressed material obscene as to minors. It did not decide whether age verification may be imposed on indecent speech that falls short of that standard, whether a covered site may be made to verify for all of its content rather than the sexual portion, or whether any particular biometric method survives review.[7] A general-audience platform is the opposite of the site *Paxton* addressed, because the overwhelming share of its content is not obscene as to minors, and for that content minors and adults hold coextensive rights of access.

2. The ceiling enforced: the post-Paxton injunction wave

The year since *Paxton* has tested exactly that boundary, and the district courts have policed it. In *NetChoice v. Miyares*, the Eastern District of Virginia preliminarily enjoined a statute requiring social-media platforms to verify every user's age and to cap minors' use at one hour per day absent parental consent. The court held the law content based because its exemptions tracked subject matter, and it concluded the scheme was likely not narrowly tailored precisely because it required all persons to verify their age before accessing speech that is protected for everyone.[8] The same decision rejected the State's parental-empowerment framing in terms that matter for any inverted design: the government may enforce a prohibition a parent has created, but it may not itself block minors' access to protected speech and then require parents to veto the government's default.[9]

---

6. See Reno v. ACLU, 521 U.S. 844, 875 (1997) (invalidating the Communications Decency Act's indecency provisions); Ashcroft v. ACLU, 542 U.S. 656 (2004) (affirming the preliminary injunction against the Child Online Protection Act on the ground that less restrictive alternatives, notably filtering, were plausibly more effective).

7. *Paxton*, 606 U.S. 461. These reservations mark the ceiling: the case that authorized age-gating did so for a defined category and left the general-audience question and the method question open. The dissent would have applied strict scrutiny to the burden on adults' access to speech that is obscene only for children. *Id.* (Kagan, J., dissenting).

8. NetChoice v. Miyares, No. 1:25-cv-02067 (E.D. Va. Feb. 27, 2026) (preliminary injunction). The court reasoned that the statute's carve-outs for news, sports, entertainment, and ecommerce content rendered it content based under Reed v. Town of Gilbert, 576 U.S. 155, 163, 170 (2015), and that conditioning access on universal age verification burdened substantially more speech than necessary.

9. *Miyares*, No. 1:25-cv-02067. The court drew the principle from Brown v. Entertainment Merchants Ass'n, 564 U.S. 786, 794–95 (2011), which reaffirmed that minors hold a significant measure of First Amendment protection and

Parallel injunctions against general-audience age-verification and design mandates have issued across the country, on materially similar reasoning.[10]

Brown v. Entertainment Merchants Association supplies the principle beneath the pattern: a State possesses legitimate power to protect children from harm, but that power does not include a free-floating authority to restrict the ideas to which children may be exposed.[11] *Paxton* is a doorway cut into that wall for one category of speech. It is not a demolition of the wall.

3. The access wall

The second, independent ground falls hardest on the population a civil-rights analysis is least free to wave off. Verification of the kind these measures contemplate depends in practice on a government-issued credential, and a substantial share of citizens do not hold one. A national analysis found that roughly 2.6 million adult citizens have no government-issued photo identification at all, and that the shortfall concentrates among disabled, elderly, and low-income Americans; about twenty percent of adults who identify as disabled lack a current driver's license, against six percent of non-disabled adults.[12] An identification requirement is therefore not a neutral formality. It is a differential barrier that tracks disability and poverty, and the circularity of the acquisition process, in which identification is needed to obtain identification, compounds the burden for exactly the people least able to absorb it. A universal gate converts that differential burden into a condition on reaching the entire lawful internet. A design confined to a narrow tier converts it into a condition on reaching one small category, which Part IV takes up.

---

that speech that is neither obscene as to youths nor otherwise proscribable cannot be suppressed solely to shield them from ideas a legislature deems unsuitable, quoting Erznoznik v. City of Jacksonville, 422 U.S. 205, 212–14 (1975).

10. See, e.g., NetChoice v. Griffin, No. 5:25-cv-05140 (W.D. Ark. Dec. 15, 2025) (enjoining a platform-design liability statute as content based and unconstitutionally vague); NetChoice, LLC v. Yost, No. 2:24-cv-00047 (S.D. Ohio Apr. 16, 2025) (permanently enjoining a parental-consent social-media access statute). The pattern is uniform enough that the constitutional question for general-audience gating can no longer be called open at the district level.

11. *Brown*, 564 U.S. at 794. The Ninth Circuit invoked the same principle in striking down California's design-code mandate. Netchoice, LLC v. Bonta, 113 F.4th 1101 (9th Cir. 2024), discussed *infra* Part IV.B.

12. CTR. FOR DEMOCRACY & CIVIC ENGAGEMENT, UNIV. OF MD., *Who Lacks ID in America Today?* (2024), `https://cdce.umd.edu/` (last visited July 5, 2026); see also VOTERIDERS, *Who Didn't Have Photo ID in 2020?* (2023), `https://voteriders.org/article/why-americans-dont-have-id/` (last visited July 5, 2026).

B. Government Prominence Mandates and the Rebalancing Bar

If verifying the audience is one wrong answer, ordering the platform to elevate approved speakers is the other. The British proposal that opened this discussion would require video-sharing services to give greater prominence to designated public-service broadcasters.[13] Whatever its merits under a legal order without a First Amendment, the analogous mandate could not stand in the United States.

Moody v. NetChoice, LLC supplies the reason. The Court held that an entity compiling and curating others' speech into an expressive product of its own receives First Amendment protection when directed to accommodate messages it would prefer to exclude, and that a platform's choices about which messages are appropriate for its feeds constitute the exercise of protected editorial control, algorithms included.[14] The Court then drew the line a prominence mandate crosses: a State's asserted interest in correcting the mix of viewpoints a platform presents relates to the suppression of expression and is not valid, because a State may not interfere with private actors' speech to advance its own vision of ideological balance, and it may not prohibit or compel speech to rebalance the speech market.[15] A rule requiring a platform to surface government-designated news over whatever its recommendation system would otherwise present is a rebalancing of the speech market by the State, the precise object *Moody* forecloses. The prominence answer thus fails on the platform's rights, just as the verify-everyone answer fails on the audience's.

---

13. See DEP'T FOR CULTURE, MEDIA & SPORT, *supra* note 1.

14. Moody v. NetChoice, LLC, 603 U.S. 707, 731, 740 (2024). The Court placed the platforms in the line of the newspaper editor in Miami Herald Publishing Co. v. Tornillo, 418 U.S. 241, 258 (1974), and the parade organizer in Hurley v. Irish-American Gay, Lesbian & Bisexual Group of Boston, Inc., 515 U.S. 557, 569–74 (1995), and it emphasized that the protection does not diminish merely because a compiler includes most submissions and excludes only the handful it most disfavors.

15. *Moody*, 603 U.S. 707. The Court added that the First Amendment achieves a diverse expressive realm by preventing the government from tilting public debate in a preferred direction rather than by licensing the government to prefer some private speakers over others.

III. The Symmetry Principle: The Feed as the Platform's Own Speech

Between the two failed answers sits the platform itself, and the doctrine of the last two years has quietly built the engine a distributor duty can run on. The engine is a symmetry.

A. Moody: Curation as Protected Expression

The first half of the symmetry is the holding just described. When a platform decides which third-party content its feed will display and how the display will be ordered and organized, it is making expressive choices; the curated feed is the platform's own expressive product.[16] That is the shield. It is what defeats the prominence mandate, and it is what the platforms invoke whenever a government reaches for their recommendation systems.

B. Anderson: Curation as Accountable Conduct

The second half arrived less than two months later, and it is the same proposition wearing the other face. In *Anderson v. TikTok, Inc.*, the Third Circuit confronted the platform's algorithmic recommendation of a "Blackout Challenge" video to a ten-year-old, who imitated it and died. Section 230, the court held, immunizes an interactive computer service only when it is sued for someone else's expressive activity, third-party speech; it does not immunize the service for its own expressive activity, first-party speech.[17] And the court drew the first-party conclusion directly from *Moody*: given the Supreme Court's observation that platforms engage in protected first-party speech when they curate compilations of others' content through their expressive algorithms, it follows that doing so amounts to first-party speech under Section 230 as well.[18]

---

16. *Moody*, 603 U.S. at 740. The proposition holds even where the curation is executed by algorithms implementing the platform's standards and guidelines.

17. Anderson v. TikTok, Inc., 116 F.4th 180 (3d Cir. 2024). Section 230(c)(1) provides that no provider of an interactive computer service shall be treated as the publisher or speaker of information provided by another information content provider. 47 U.S.C. §230(c)(1).

18. *Anderson*, 116 F.4th at 183–84. Because the recommendation delivered through the platform's curated feed was the platform's own expressive activity, and because that recommendation formed the basis of the plaintiff's claims, Section 230 did not bar the suit. The holding is limited to affirmative promotion through a curated algorithm; the court acknowledged that Section 230 still precludes liability for the mere presence of third-party videos on the platform.

The symmetry is the point. A platform cannot coherently claim the feed as its own expression when the question is constitutional protection and disclaim the feed as someone else's speech when the question is accountability. *Moody* and *Anderson* are one proposition read in two postures, and together they locate responsibility for the curated feed, including the classification decisions that govern what the feed shows a child, in the platform's own conduct.

C. The Unsettled Edge

Candor requires marking the edge. *Anderson* sits in tension with an older line holding that Section 230 immunizes the arrangement and distribution of third-party content even through personalized algorithms,[19] and the Supreme Court in *Moody* itself reserved the hardest attribution question: whether a determination made by a machine-learning system operating without a human in the loop carries the same expressive character as a human editorial choice.[20] A distributor duty should therefore not be built to depend on *Anderson* winning everywhere. The design below does not. It regulates the platform's classification conduct directly, in a register the platform already occupies by statute, and it needs *Anderson* only for what *Anderson* uncontroversially shows: that the platform's own conduct in shaping what reaches children is a legally cognizable object of responsibility rather than a void Section 230 has sealed off.

IV. The Distributor Verification Duty

The proposal follows from the space the doctrine leaves. Put the obligation on the party that decides and profits, require it to verify its own classifications rather than trust the uploader, draft the duty inside the compelled-speech and vagueness constraints the courts have marked, scale it to the platform's capacity, and gate the narrow adult tier with proof that reveals nothing.

---

19. See Force v. Facebook, Inc., 934 F.3d 53 (2d Cir. 2019) (holding that algorithmic matching of users to third-party content did not strip Section 230 immunity, over a partial dissent). District courts continue to collect cases on both sides. The disagreement is a live circuit-level tension rather than a settled rule.

20. *Moody*, 603 U.S. 707 (Barrett, J., concurring) (contrasting an algorithm that implements a human editorial choice, which the First Amendment protects, with an automated system that merely serves engagement predictions, whose First Amendment status may differ).

A. The Self-Report Sieve and a Capability That Already Exists

The objection that a platform cannot be expected to inspect its own catalog at scale is answered by the platform's existing conduct. Video services already scan entire uploads against a content taxonomy for a commercial purpose, applying advertiser-suitability rules that turn on the presence and density of disfavored material across a whole video rather than a sampled fragment.[21] The whole-video classification the platform declines to build for child safety is one it has already built for advertiser safety. The gap is one of priority rather than capability: the same pipeline that protects a brand can be pointed at the protection of a child. Meanwhile the child-facing classification, the "made for kids" designation, continues to rest in the first instance on the uploader's checkbox, with automated review as a backstop.[22] A verification duty makes the priority a legal one and shifts the classification from a declaration the uploader controls to a determination the platform must stand behind.

B. Drafting Inside Bonta and Griffin: Facts, Not Opinions; Standards, Not Sensibilities

Two recent decisions teach, with unusual precision, how such a duty must and must not be written.

The first is *NetChoice, LLC v. Bonta*. California's Age-Appropriate Design Code required covered businesses to produce reports opining on whether their services' designs could expose children to harmful or potentially harmful content, and to mitigate the risks they identified. The Ninth Circuit held the requirement facially unconstitutional: it compelled speech by forcing businesses to opine on potential speech-based harms to children, and because the compelled opinions concerned controversial issues of public concern rather than purely factual commercial disclosures, the requirement drew strict scrutiny it could not survive.[23] The lesson is not that

---

21. See YOUTUBE, *Advertiser-Friendly Content Guidelines*, `https://support.google.com/youtube/answer/6162278` (last visited July 5, 2026); YOUTUBE, *Inappropriate Language: Advertiser-Friendly Guidelines*, `https://support.google.com/youtube/answer/9725604` (last visited July 5, 2026) (rating profanity by placement and frequency across the full runtime, with distinct treatment of the opening seconds).

22. See *supra* note 2. The designation exists to trigger COPPA's data-collection limits, and the platform's enforcement posture around it reflects that privacy origin rather than a content-safety verification mandate.

23. Netchoice, LLC v. Bonta, 113 F.4th 1101 (9th Cir. 2024). The court reasoned that the creation of information is speech, that the right to refrain from speaking is protected, and that a mandate to identify and mitigate "harmful

platform duties are impossible; the court left open whether a more limited requirement focused on factual matters such as data practices, without content-based risk opinion, would survive.[24] The lesson is that the duty must attach to a factual determination, and the determination this paper's duty attaches to is one federal law already obliges the platform to make: whether a given upload is child-directed within the meaning of the COPPA rule. Verifying the accuracy of a classification the law already requires is an evidentiary obligation about a fact, an audit, and stands on different footing from compelled opinion about contested harm.

The second is *NetChoice v. Griffin*. Arkansas imposed liability on platforms whose designs caused users to develop harms the statute described, keyed to what the platform "should have known" about a user's sensitivities. The court found the scheme unconstitutionally vague because it failed to specify a standard of conduct to which platforms could conform, leaving violation to depend on the sensibilities of an unspecified user and a jury's retrospective judgment.[25] The teaching is the mirror image of *Bonta*'s: the duty must run to a specified, pre-existing classification standard, applied prospectively by the platform itself, rather than to an open-ended harm judgment reconstructed after the fact. A duty to verify child-directed designations against the COPPA rule's own factors is exactly that.

C. Scaling the Duty to Capacity

A duty stated as a flat command to inspect every upload would misfire in a familiar way. A universal scanning mandate is a rounding error for the largest incumbent and an existential cost for a small competitor or a nonprofit host, so a careless rule would entrench the very firm it means to discipline. The duty-of-care architecture emerging on both sides of the Atlantic already

---

or potentially harmful" content forces a subjective, content-based judgment far removed from the purely factual disclosures tolerated under Zauderer v. Office of Disciplinary Counsel, 471 U.S. 626 (1985).

24. *Bonta*, 113 F.4th 1101. The opinion expressly declined to decide the constitutionality of a narrower assessment obligation confined to system design and data-management facts.

25. NetChoice v. Griffin, No. 5:25-cv-05140 (W.D. Ark. Dec. 15, 2025). The court also confirmed that the core of Section 230 protects a platform's moderation decisions about which third-party content to remove and which to permit, a point consistent with the design below, which compels no removal and no carriage; accord Moody v. NetChoice, LLC, 603 U.S. 707, 718 (2024).

calibrates obligations to a service's size and risk rather than imposing a single fixed command,[26] and the verification duty should adopt the same posture: whole-catalog verification from the global platform that has already demonstrated the capability, proportionate diligence from the small host, with the threshold indexed to capacity and revenue rather than to a nominal user count a startup can cross before it can afford the machinery.

D. Age Assurance Without Surveillance

The narrow adult tier still needs a gate, and the gate should prove a fact without harvesting one. The design question is now the subject of a converging technical literature: the credible architectures divide into direct submission of identity documents, verifiable digital credentials presented from a wallet, and biometric age estimation, and they differ radically in what they disclose and to whom.[27]

1. Zero-Knowledge Age Tokens

The privacy-preserving option is in market and in regulatory prototype. A digital wallet can hold a government-issued credential and, using a zero-knowledge proof, return to a requesting service a single bit, whether the holder meets an age threshold, without transmitting a birthdate, a name, or a document.[28] The French data-protection authority's innovation laboratory built a working demonstrator of precisely this architecture, proving an age through a certified third party while guaranteeing data minimization and protecting the user's identity from the requesting

---

26. See Online Safety Act 2023, c. 50 (UK) (imposing risk-assessment and safety duties differentiated by service category and reach); Kids Online Safety Act, S. 1409, 118th Cong. (2023) (proposing a duty-of-care model for covered platforms). The point here is architectural rather than an endorsement of any particular bill's scope.

27. See Shuang Liu & Sarah Scheffler, *Adequately Tailoring Age Verification Regulations* (arXiv:2601.20241, 2026) (mapping the post-*Paxton* legislative landscape and evaluating verification architectures against an adequate-tailoring model); Wojciech Wodo, Maksymilian Górski & Lucjan Hanzlik, *Age Verification in the Web: Holy Grail to Control Access to Restricted Content* (arXiv:2605.05513, 2026) (analyzing deployed and proposed European age-verification systems and their privacy properties).

28. See GOOGLE, *New Age and Identity Verification Features in Google Wallet* (2025), `https://blog.google/products/google-pay/google-wallet-age-identity-verifications/` (last visited July 5, 2026). Zero-knowledge proofs allow a prover to demonstrate the truth of a statement, here, that a signed credential's birthdate lies at least eighteen years in the past, without disclosing anything beyond the statement's truth. See Bjorn Oude Roelink, Mohammed El-Hajj & Dipti Kapoor Sarmah, *Systematic Review: Comparing zk-SNARK, zk-STARK, and Bulletproof Protocols for Privacy-Preserving Authentication*, 7 SECURITY & PRIVACY e401 (2024) (surveying the protocol families and noting age verification as a principal application).

site,[29] and the European Union's harmonized age-verification application is being built on the same wallet infrastructure with zero-knowledge integration planned.[30] This is the opposite of the upload-your-identification model, which manufactures a store of sensitive records at every gate and rewards the platform with data it can monetize or lose. A design that solves a child-safety problem by enlarging a surveillance dividend has answered the wrong question; the zero-knowledge token answers the right one by collecting nothing.

2. Why Facial Estimation Fails

The tempting fallback, estimating age from a live selfie so that no credential is needed, cannot carry the gate, and the reason is empirical on two independent axes. The first is accuracy at the boundary. Facial age estimation is a regression with an error band of years: the published benchmarks that defined the field report mean absolute errors ranging from roughly three to six years for algorithmic estimators, with human observers performing no better,[31] and the one line that matters, seventeen against eighteen, sits inside that band. The second is adversarial robustness in deployment. The Internet Matters field study found children defeating deployed estimation systems with a drawn moustache, with exaggerated expressions, and by pointing the camera at adult-looking video-game characters.[32] A gate is only as strong as its easiest entrance. A system that admits a strong credential and a weak selfie is a system whose real security is the selfie, because the weak path becomes the default attack surface for every user the gate exists to stop.

---

29. OECD, *Emerging Privacy-Enhancing Technologies: Current Regulatory and Policy Approaches* (2023), `https://doi.org/10.1787/bf121be4-en` (last visited July 5, 2026) (describing the CNIL innovation laboratory's zero-knowledge age-verification demonstrator and its three-party architecture of service provider, certified verifier, and certifying authority).

30. See Wodo et al., *supra* note 27 (describing the European age-verification solution built on the European Digital Identity Wallet and its planned integration of zero-knowledge proofs).

31. See Gabriel Panis, Andreas Lanitis, Nicolas Tsapatsoulis & Timothy F. Cootes, *Overview of Research on Facial Ageing Using the FG-NET Ageing Database*, 5 IET BIOMETRICS 37 (2016) (compiling benchmark results, with the strongest reported algorithm at a mean absolute error of 3.17 years and typical methods between 4.4 and 6.2 years, against a human-observer benchmark of 4.7 years).

32. INTERNET MATTERS, *supra* note 3.

3. Why the Social Security Number Is the Wrong Credential

The instinct to widen coverage by accepting a Social Security number instead of a photo credential fails at the mechanism. A zero-knowledge age proof must be generated from a credential that contains a signed date of birth; the Social Security number contains none, so there is nothing in it from which to prove an age. The number was never designed as an authenticator, it is a static secret that cannot be rotated once exposed, and it has been exposed at national scale: the 2017 Equifax breach alone compromised the personal information, including Social Security numbers, of roughly 147 million people, a figure the Federal Trade Commission's settlement announcement placed at the center of the case.[33] Building a child-safety gate on the most widely breached identifier in the country would trade a narrow, reversible access gap for a broad and irreversible security exposure borne by every user, a worse allocation of harm by the measure this paper uses everywhere else.

V. Why the Design Threads Every Needle

The distributor verification duty is not a compromise that dissatisfies each constraint a little. It satisfies each one.

It clears the constitutional ceiling. Its adult tier gates only the category *Paxton* authorized and leaves the general-audience floor open to everyone, so it never asks the reserved question about indecent or general-audience speech that the post-*Paxton* injunctions have answered against the States.[34] Its platform-facing obligation regulates the provider's own classification conduct, the conduct *Moody* attributes to the platform as its own and *Anderson* holds accountable as its own, and it compels neither carriage nor elevation of any speaker, so it stays clear of the rebalancing

33. Press Release, Fed. Trade Comm'n, *Equifax to Pay $575 Million as Part of Settlement with FTC, CFPB, and States Related to 2017 Data Breach* (July 22, 2019), `https://www.ftc.gov/news-events/news/press-releases/2019/07/equifax-pay-575-million-part-settlement-ftc-cfpb-states-related-2017-data-breach` (last visited July 5, 2026). On the re-identification risk of supposedly anonymized identity attributes generally, see Liu & Scheffler, *supra* note 27.

34. Free Speech Coal., Inc. v. Paxton, 606 U.S. 461 (2025); NetChoice v. Miyares, No. 1:25-cv-02067 (E.D. Va. Feb. 27, 2026).

bar.[35] It is drafted to the surviving side of the compelled-speech line, verifying a factual statutory classification rather than compelling opinion about contested harms, and to the surviving side of the vagueness line, running to a specified pre-existing standard rather than to a jury's reconstruction of what the platform should have known.[36]

It protects privacy by construction, because the zero-knowledge token returns a bit and stores nothing, so the design creates no honeypot and confers no data dividend on the platform.[37] It contains the access cost, because only the narrow adult tier turns on a credential; the general-audience floor, where news, information, and the overwhelming bulk of ordinary speech live, stays open to the disabled, elderly, and low-income adults whom an audience-wide rule would exclude.[38] And it declines to crown the incumbent, because the duty scales to capacity rather than commanding a uniform cost that only the largest firm can bear.

VI. Limitations and Open Questions

Four limits deserve statement rather than concealment. The first is the general-audience middle. The proposal deliberately leaves the vast center of ordinary content ungated, which means the hard problem of detecting material engineered to harm children while remaining facially innocuous, the keyword-clean and filter-evasive material, is relocated rather than dissolved; detecting a profanity is a solved classification problem, and detecting engineered child-directed harm is a materially harder one, even on the same pipeline. The second is the identification residual. Privacy-preserving proof still depends on holding a credential, and although the design confines that dependency to the adult tier, the coverage gap for the credential-less is a cost the tier boundary bounds without erasing; broadening the set of accepted credentials narrows the gap without cracking the gate, and that is the correct direction of repair. The third is the Section 230 tension. *Anderson* and the older

35. Moody v. NetChoice, LLC, 603 U.S. 707 (2024); Anderson v. TikTok, Inc., 116 F.4th 180 (3d Cir. 2024).

36. Netchoice, LLC v. Bonta, 113 F.4th 1101 (9th Cir. 2024); NetChoice v. Griffin, No. 5:25-cv-05140 (W.D. Ark. Dec. 15, 2025).

37. See *supra* notes 28–30 and accompanying text.

38. See CTR. FOR DEMOCRACY & CIVIC ENGAGEMENT, *supra* note 12. The residual exclusion is real but bounded: an adult without a credential loses the adults-only slice and keeps everything else.

matching-algorithm cases have not been reconciled, and a duty built on classification conduct should expect the platforms to litigate the boundary.[39] The fourth is the attribution question *Moody* reserved: as classification shifts from human review toward machine-learning systems, the expressive character of an automated determination, and therefore its constitutional treatment, becomes less settled, and a verification duty that leans on automated classification inherits that unsettled question.[40] Enforcement design, penalty calibration, and the divergence between the United States ceiling and the more permissive posture of other jurisdictions remain open as well.[41]

VII. Conclusion

The two answers now dominating the debate share a common error of address. Verifying the audience aims child protection at the user and collides with the ceiling *Paxton* drew and the district courts have spent a year enforcing; prominencing the feed aims it at the speaker and collides with the autonomy *Moody* secured. The party that classifies, distributes, and profits sits between them, and the doctrine now speaks to that party in both registers at once: the feed is the platform's own expression when it seeks protection, and the platform's own conduct when the question is responsibility. Placing the burden there, as a verification duty scaled to capacity, drafted to factual classification standards, and paired with a gate that proves an age while collecting nothing, is the allocation the doctrine already permits and the record already recommends. A platform that will scan every second of a video to protect an advertiser can be required to do the same to protect a child.

39. Compare Anderson v. TikTok, Inc., 116 F.4th 180 (3d Cir. 2024), with Force v. Facebook, Inc., 934 F.3d 53 (2d Cir. 2019).

40. Moody v. NetChoice, LLC, 603 U.S. 707 (2024) (Barrett, J., concurring).

41. On the divergence, compare the architecture of the Online Safety Act 2023, c. 50 (UK), with the injunction pattern described *supra* Part II.A.2. For the technical-policy synthesis of that divergence, see Liu & Scheffler, *supra* note 27; Wodo et al., *supra* note 27.