PUBLIC KNOWLEDGE

GETTING AGE ASSURANCE RIGHT: A RISK-BASED FRAMEWORK FOR HIGH-RISK ONLINE FEATURES

2026 | MAY

SARA COLLINS

Getting Age Assurance Right: A Risk-Based Framework for High-Risk Online Features

Sara Collins

Table of Contents

I. Introduction **4**
II. Defining "High-Risk": Why Feature-Level Classification Matters **5**
III. A Three-Tier Risk Framework **7**
Tier 1 — High Risk 7
Tier 2 — Medium Risk 8
Tier 3 — Low Risk 9
IV. The Drop-Off Problem: Why Method Choice Is a Safety Question **10**
V. Acceptable Methods by Tier **12**
Why Government ID Is Excluded as a Required Method 12
Minimum Floor for Any Acceptable Method: FTC Safe Harbor Conditions 13
Acceptable Tier 1 Methods 14
Acceptable Tier 2 Methods 15
VI. The Knowledge Asymmetry Rule **16**
VII. Vendor Transparency and Demographic Accuracy **17**
VIII. Conclusion **18**

I. Introduction

In "The Kids Aren't Alright Online: How to Build a Safer, Better Internet for Everyone," Public Knowledge argued that age assurance – the practice of verifying an online user's age – should be a narrow tool, not a policy solution. The most important thing lawmakers can do for children online is require technology companies to design safer products — not deputize parents and children to police systems that were never built with kids' wellbeing in mind in the first place. But that argument, intentionally, left a question open: when age assurance is appropriate, how should it work?

The policy moment demands an answer. More than 25 states have now passed some form of age assurance requirement, and the legislative patchwork is still growing.[1] At the federal level, the Federal Trade Commission issued an Enforcement Policy Statement in 2026[2] promoting the adoption of age-verification technology under the Children's Online Privacy Protection Act, or COPPA, rule, marking the first time a federal agency has affirmatively incentivized specific age-verification approaches through the promise of enforcement forbearance, or the temporary suspension of enforcement actions by an agency. The FTC's guidance is significant not just for what it says, but for what it leaves unsaid.

Three failures have defined the current landscape. First, no agreed definition exists of which features or products actually require age assurance, leaving companies and regulators to haggle over vague standards like "reasonable efforts" and "commercially reasonable." Second, no agreed definition exists of what methods are acceptable, meaning the market for age assurance technology operates without meaningful quality floors or accountability mechanisms. Third, and most critically, almost no regulatory framework accounts for what happens when verification fails not by letting the wrong users in, but by locking the right users out. This is the "drop-off" problem, and it is not a footnote to the age assurance debate. Running beneath all three failures is a more fundamental problem: Policymakers have aimed age assurance at the wrong target. Blocking children from "harmful" content, the goal that has animated most age verification legislation to date, is a distraction from the design-level interventions that can actually protect children online and better survive First Amendment scrutiny. Age assurance, when it is used at all, should be understood as a tool for delivering safer default experiences to users who may be minors; not as a content firewall.

This paper proposes a risk-tiered framework that addresses all three failures. It matches the required strength of age assurance to the actual harm potential of a given feature; privileges privacy-preserving methods over identity-exposing ones; categorically excludes government ID

[1] Free Speech Coalition, *State Age Verification Laws*, Action Center (last visited Apr. 3, 2026), https://action.freespeechcoalition.com/age-verification-resources/state-avs-laws/. The listed states with laws currently in effect are: Alabama, Arizona, Arkansas, Florida, Georgia, Idaho, Indiana, Kansas, Kentucky, Louisiana, Mississippi, Missouri, Montana, Nebraska, North Carolina, North Dakota, Ohio, Oklahoma, South Carolina, South Dakota, Tennessee, Texas, Utah, Virginia, and Wyoming.
[2] Federal Trade Commission, *Enforcement Policy Statement Promoting the Adoption of Age-Verification Technology* (Feb. 25, 2026), https://www.ftc.gov/legal-library/browse/enforcement-policy-statement-promoting-adoption-age-verification-technology.

as a required verification method; and builds the drop-off problem into the structure of the framework rather than ignoring it. Critically, it reorients the purpose of age assurance away from content blocking and toward what age signals can actually accomplish: calibrating a user's experience on a platform to account for their likely age. The goal is a framework that meaningfully adjusts how platforms behave for users who may be minors.

II. Defining "High-Risk": Why Feature-Level Classification Matters

Feature-level classification matters because the platform itself is the wrong unit of analysis. The same service can host features that pose genuinely severe risks to minors alongside features that are entirely benign. An app that allows users to browse public videos; message strangers privately; stream live video to unknown audiences; and post comments in a closed friend group is doing four quite different things, each with a different risk profile from a child safety standpoint. A law that treats the whole platform as either requiring age assurance or not will either dramatically over-restrict (imposing heavy verification burdens on access to low-risk features) or dramatically under-restrict (permitting the most harmful features to continue operating without any meaningful check). The right unit of analysis is the *feature*, evaluated along two axes: the probability that minors will encounter harm through that feature, and the severity of the harm if they do.[3]

Risk assessment must also account for default settings. An open direct messaging feature that allows any unknown user to initiate contact with any other user poses a categorically different risk than a messaging feature that is opt-in and limited to established connections. The same logic applies to live video, location sharing, and algorithmic recommendations. Assessing risk at the feature level means assessing it in the context of how the feature actually operates by default; not in the best-case configuration that a careful user might eventually choose. Once again, the goal is not to identify and block every minor, but to ensure that the default experience a platform delivers is appropriate for the realistic range of users who will encounter it.

There is also a technical dimension to age-appropriate experiences that the policy literature has not fully absorbed. Features that are blocked from being used by minors entirely impose different requirements on age assurance systems than features designed to deliver safer defaults to users who *may be* minors. For the former, the system must actually exclude — which means it must be resistant to motivated circumvention by users who know they are being blocked. For the latter, the system need only provide a reasonable signal about likely age to adjust the user's experience, which is a much less demanding task. Conflating these two use

[3] Organisation for Economic Co-operation and Development, DSTI/CDEP/DGP(2020)3, Children in the Digital Environment: Revised Typology of Risks, Unclassified (Jan. 12, 2021). This approach draws on the OECD's typology of online harms to children, which identifies four primary harm categories — content harms (exposure to harmful material), contact harms (unwanted or dangerous communication with others), conduct harms (being manipulated or exploited through platform mechanics), and commercial harms (targeting or exploitation through data collection and advertising).

cases produces frameworks that are simultaneously too demanding for low-risk features and not demanding enough in specifying what "blocking" actually requires.[4]

The three-tier framework proposed addresses this directly. Each tier corresponds to a different risk level, a different default treatment, and a different set of acceptable assurance methods — all derived from what the technical literature tells us age assurance systems can actually do.

III. A Three-Tier Risk Framework

Tier 1 — High Risk

The highest-risk features are those where a minor's exposure, if it occurs, is likely to cause serious and potentially irreversible harm, and where the feature itself creates the conditions for that harm. The features that fall into this high-risk tier share a structural property: they allow encounters between users who have not established any prior connection, and the harm potential flows directly from that openness. Open live video streaming to unknown audiences is another. Unsolicited direct messages from strangers, the mechanism through which online predation most commonly begins, belong here as well.[5] So does public live location mapping of the kind that allows any user to see another user's real-time location, a feature whose risk to minors has been documented across multiple platforms.[6] What these high-risk features have in common is a structural design choice that places minors in proximity to unknown adults.

---

[4] *See* Eric Rescorla et al., *Age Assurance Online: A Technical Assessment of Current Systems and Their Limitations*, Knight Georgetown Institute (2024) [hereinafter *KGI Technical Assessment*], at iv (identifying two primary and technically distinct use cases — "safer defaults" and "blocking" — and finding that "[t]he suitability of different age assurance mechanisms varies significantly depending on whether the goal is to provide safer defaults or to block access entirely"); *id.* at 6–7 (explaining that blocking use cases involve anonymous or pseudonymous users with no prior service relationship and greater circumvention motivation, imposing categorically different technical requirements than safer-defaults contexts).
[5] *See* Complaint for Abatement and Civil Penalties, *State of New Mexico ex rel. Torrez v. Meta Platforms, Inc.*, No. D-101-CV-2023-02838 (1st Jud. Dist. Ct., Santa Fe Cty., Dec. 5, 2023).
[6] *See* Social Media Victims Law Center, *Snapchat Fentanyl Lawsuits*, https://socialmediavictims.org/snapchat-lawsuit/fentanyl/ (last visited Apr. 3, 2026).

It is worth acknowledging that the Supreme Court's 2025 decision in *Free Speech Coalition v. Paxton* has animated much of the current legislative activity around age assurance. The decision confirmed that age verification requirements for access to explicit sexual content can survive First Amendment scrutiny under certain conditions. Legislators have taken this decision as a green light to implement age verification broadly. But the technical barriers to blocking determined users are significant; the harms to legitimate adult users from a privacy and access information standpoint are many; and the design-level interventions that most reliably protect children operate upstream of any individual content encounter. The framework proposed here focuses on what age assurance can actually accomplish on general-purpose platforms. Content blocking, even where constitutionally permitted, is not the right organizing principle.

For Tier 1 features, the default is denial of access absent a meaningful age signal. Importantly, that signal must not include presenting a government ID, which provides the user's identity as a byproduct of age verification.

Tier 2 — Medium Risk

The second tier covers features that pose real risks to minors but where the mechanism of harm is different — typically operating through design choices that exploit psychological vulnerabilities over time rather than through a single harmful encounter. Push notifications engineered to maximize re-engagement; autoplay and endless scroll mechanics that eliminate natural stopping points; gambling-like reward mechanics like loot boxes in gaming environments; and behaviorally targeted advertising directed at users who may be minors all belong in this category.[7]

---

[7] *See* Lisa Macpherson & Morgan Wilsmann, *A Policy Primer for Free Expression and Content Moderation, Part III: Safeguarding Users*, Public Knowledge (Dec. 9, 2024), https://publicknowledge.org/safeguarding-users/ (describing how the advertising-driven business model incents platforms to design features that "maximize the time and energy people put into searching, scrolling, liking, commenting, and viewing" and has been linked to "compulsive use, unhealthy behaviors, social isolation, and unsafe connections"). *See Also* Sara Collins & Morgan Wilsmann, *The Kids Aren't Alright Online,* Public Knowledge, 4, 22–23 (Aug. 2025) (observing that teens themselves describe losing hours to algorithmic feeds engineered to prevent disengagement, and that "the knowledge of this practice

These features tend to have a different structural property than Tier 1: Medium-risk features typically involve *ongoing service relationships*, which means that behavioral and account-based signals — useless for first-contact verification — become viable for age assessment. A user who has been on a platform for three months has generated a behavioral profile that can reasonably be assessed for age-consistent patterns. That is not a substitute for harder verification where the risk of harm is high, but it is a meaningful signal for ongoing feature-level access decisions.

The circumvention calculus also differs. A minor who is determined to circumvent a content block has a strong incentive to do so. Designing policy around that use case leads to frameworks that impose severe costs while delivering little safety benefit. A minor who is simply trying to use a social media app and encounters safer-default settings, quieter notifications, no autoplay, no targeted ads, has far less incentive to go to elaborate lengths to circumvent anything. This difference in motivation has direct implications for how much friction is actually needed. Mandating government ID verification for access to push notification settings would impose severe costs to privacy and free expression, while delivering almost no additional safety benefit for minor users.

For Tier 2 features, the default is restricted access with graduated pathways to fuller access based on age signals proportionate to the risk.

Tier 3 — Low Risk

The third tier covers features where the protective work can be done entirely through design rather than age gating. Basic content browsing; access to educational resources; search; participation in closed communities where membership is controlled; and standard interaction features like posting, commenting, and liking, when paired with robust content moderation and user safety tools, do not require age gates. *They require good design*.

---

does not necessarily cause them to log off," illustrating that awareness of manipulation does not neutralize its effect).

This is not a concession to risk. It is a recognition that age assurance imposes costs to privacy, to access, and to the chilling of legitimate use, and that those costs are only justified when they are proportionate to the harm they prevent. Imposing age gates on low-risk features does not make children safer. It makes the internet less useful for everyone while creating a false sense of security that distracts from the design-level interventions that actually matter.

For Tier 3 low-risk features, safe-by-design defaults are the appropriate and sufficient standard. No additional data collection for age determination is warranted, and no age gate should be required.

IV. The Drop-Off Problem: Why Method Choice Is a Safety Question

Age assurance is most commonly evaluated as a question about minors: Does it work? Does it keep underage users out? These are legitimate questions, but they are not the *only* questions, and a framework that answers only them is incomplete as a matter of both policy *and* ethics.

The other question is whether age assurance works for adults: Does it let legitimate users in? The evidence on this point is stark, and it has received almost no attention in the legislative process. Research from Carnegie Mellon University found that government ID verification was completed by only 22 to 28 percent of adult users in a controlled study. Adding a liveness check,[8] like blinking during a selfie scan, dropped completion to 18 percent. AI facial estimation fared better, at 51.5 percent, but still excluded roughly half of adult users. Email-based verification achieved the highest completion rate, at 86 percent, though it also provides the weakest age signal.[9] Notably, even among users who did complete the verification process, a majority reported being "somewhat uncomfortable" with every non-checkbox method. Discomfort matters even when it does not produce non-completion, because discomfort causes a chilling effect on speech and access.

At-risk populations most harmed by drop-off are users without government ID or a qualifying credit card; LGBTQ+ youth; immigrants; and domestic abuse survivors who rely on anonymity for safety.[10] Credit cards are used as an age signal in some jurisdictions because issuance is restricted to adults by law, meaning possession of a card serves as a proxy for being over 18. Where credit card verification is deployed, adults without a card are effectively locked out. According to FDIC data, 23.6 percent of U.S. households do not have a credit card, and that gap is not random: it falls disproportionately on lower-income adults, people of color, younger

[8] A secondary verification step that confirms the person submitting an identification document is physically present and is the same individual depicted in it, rather than holding up a photograph or other static image.
[9] Yanzi Lin, et al. User (Non-)Compliance with Age Verification: Preliminary Evidence from a Deceptive Web Experiment. Carnegie Mellon University CyLab Security and Privacy Institute. (2026, January). https://www.cs.cmu.edu/~sscheffl/docs/2026/AgeVerif2026.pdf *(preliminary, peer review pending)*.
[10] Sarah Forland, et al. *Age Verification: The Complicated Effort to Protect Youth Online*, New America (Apr. 23, 2024), [hereinafter *New America Age Verification Report*] https://www.newamerica.org/oti/reports/age-verification-the-complicated-effort-to-protect-youth-online/.

adults, and those who rely on cash or prepaid debit.[11] These are, in many cases, the same populations who stand to benefit most from legal access to age-restricted resources. The same is true of adults without qualifying government photo ID, of whom there are millions in the United States.

The FTC's Enforcement Policy Statement is largely silent on trade-offs in high-friction verification. Its safe harbor conditions; data minimization; prompt deletion of verification data; no secondary use; and written vendor assurances are all sound privacy protections. But they address *none* of the access harm caused by verification friction itself. The Statement creates incentives for operators to adopt more robust verification methods without acknowledging that robustness and availability exist in fundamental tension: the methods with the highest accuracy tend to have the lowest completion rates and the narrowest population coverage. A framework that ignores drop-off is incomplete as a child safety framework, because the adults most likely to be deterred by high-friction methods are disproportionately those the government has the greatest obligation to protect.

There is a rights-based dimension to this problem that is independent of completion rates. Even if a high-friction method achieved universal completion, the choice of method would still matter, because different methods expose different amounts of personal information, and users have a legitimate interest in controlling how much of their identity they must surrender in order to access a feature on a general-purpose platform. Users who are willing to confirm that they are over 18 through a zero-knowledge proof are not making the same privacy decision – or providing the same level of personally identifying information – as users who upload a passport scan to a third-party verification provider. The difference is not cosmetic. It is the difference between revealing *age eligibility* and revealing *identity*. Any age assurance framework that offers users only one method forecloses a privacy choice that users are entitled to make. The drop-off problem is real and urgent, but user choice over method is not merely a remedy for drop-off. It is a *requirement* in its own right.

The principal circumvention vectors, including VPNs deployed against server-based systems; unlocked or modified devices deployed against device-based systems; adult-minor cooperation that exploits the verification of a consenting adult on behalf of a minor; and injection attacks or deepfakes deployed against facial estimation systems are not individually closable without costs that exceed the benefit.[12] The adult-minor cooperation vector deserves particular attention. This is because most age assurance systems verify a credential (a saved session, an unlocked device, a credit card number) — *not* the person holding the credential at any given moment. A parent who verifies once effectively authorizes anyone who later reaches that credential, long

[11] Federal Deposit Insurance Corporation (FDIC), *2023 FDIC National Survey of Unbanked and Underbanked Households* 52–54 (Nov. 2024) (reporting that 76.4% of U.S. households had a traditional credit card in 2023, meaning 23.6% did not; and finding that traditional credit card ownership was substantially lower among lower-income households — with only 38.2% of households earning less than $15,000 holding a card, compared with 90.4% of households earning $75,000 or more — as well as among Black households (59.4%), Hispanic households (65.0%), and American Indian or Alaska Native households (57.0%), compared with 81.9% of White households; the youngest households (ages 15–24) also had the lowest ownership rate among non-elderly age groups at 65.8%).

[12] *KGI Technical Assessment*, at 20.

after the adult has left the room. Closing this gap would require confirming that the verified person is physically present at each session, which means biometric re-authentication on every visit, transforming every age-gated site into a recurring identity checkpoint for every adult user, indefinitely. No serious proposal has contemplated this, and none should. Age assurance creates friction, not a wall. Policymakers who draft requirements without acknowledging this structural property are setting up enforcement regimes that will disappoint them. Matching method to purpose is not just a technical consideration; it is the difference between a policy that works and one that doesn't.

V. Acceptable Methods by Tier

Before turning to which methods of age assurance are acceptable at each tier, the paper must state a prior principle: No compliant age assurance system may offer users only one method. This is not simply a practical accommodation for users who cannot access a particular method. It is an affirmative design requirement. Age assurance necessarily involves a trade-off between the strength of the age signal and the amount of personal information the user must expose to produce it. Different users will reasonably weigh that trade-off differently, and a system that eliminates the choice has made that decision for them. Platforms must offer at least two meaningfully distinct methods at each risk tier, where "meaningfully distinct" means that the methods differ in their privacy properties, not merely in their user interface.This requirement to provide genuine alternatives applies regardless of whether the primary method achieves high completion rates in aggregate. *The right to choose how one's identity is handled does not disappear because most users are willing to waive it.*

Why Government ID Is Excluded as a Required Method

Government ID (the driver's license scan, the passport upload, the national identity document check) is the method that most intuitively reads as rigorous, and it is the method that most legislative drafters reach for when they want to appear serious about age verification. It should be categorically excluded as a required Tier 1 age-assurance method.

The objections are multiple and mutually reinforcing. At the technical level, age verification using government ID is, in practice, identity verification. While conceptual frameworks like those from the National Institute of Standards and Technology and the Age Verification Providers Association maintain a formal distinction between verifying *age* and verifying *identity*, current commercial systems collapse that distinction in operation: They require the user to produce a document that establishes their full legal identity, and the evaluating system learns that identity as a byproduct of confirming the user's age. The privacy cost of this is uniquely severe: Unlike facial estimation, which reveals only a face, or behavioral inference, which reveals only behavioral patterns, government ID verification reveals the user's full legal name, date of birth, address, and any other information contained in the document.

At the structural level, government ID excludes significant portions of the adult population by definition. Teenagers between the ages of 14 and 16 rarely hold government ID, which creates an ironic problem for any system trying to determine whether a user is 16 or 17 as opposed to

13 or 14. Undocumented users lack a government ID by definition. Unbanked users frequently lack the forms of financial identification that serve as government ID proxies, like credit cards. There are also users, like journalists or activists facing government repression, who rely on their recognized constitutional right to use platforms anonymously.[13] They are directly harmed by any requirement that tethers online access to real-world identity.

The empirical record on completion rates confirms these structural exclusions in practice. Government ID verification produced the lowest completion rates and highest discomfort levels of any method tested in the CMU study, and crucially, providing reassurance language about data handling did not significantly improve completion among non-completers. The problem is not that users distrust the systems, though many do; the problem is that many users *cannot use them*.

The FTC's own Enforcement Statement implicitly acknowledges this. Rather than endorsing government ID, the Statement refers neutrally to "age estimation tools," "age-verification tools," and "age inference tools." France's data protection authority, the Commission Nationale de l'Informatique et des Libertés or CNIL, reached a similar conclusion through a different route: after a comprehensive review of available age assurance mechanisms, the CNIL concluded that no existing solution satisfactorily meets the combined requirements of verification reliability, population coverage, and privacy protection simultaneously.[14]

Furthermore, the prevailing implementation model requires users to submit government ID directly to individual platforms, with no technical constraint on retention, secondary use, or cross-platform linkage. Requiring each website or application to independently collect, validate, and store a government-issued credential creates *dozens* of new custodians of some of the most sensitive identity data a person possesses, each one a potential breach, a potential surveillance vector, and a potential commercial actor with incentives misaligned from the user's privacy interests.[15]

Minimum Floor for Any Acceptable Method: FTC Safe Harbor Conditions

Any method used for Tier 1 or Tier 2 age assurance must meet a set of baseline conditions before it qualifies as acceptable under this framework. The FTC's Enforcement Statement identifies these conditions as the basis for enforcement forbearance, and this paper adopts them as the minimum floor for any acceptable method. They are: The age-verification data may

[13] Rindala Alajaji, *10 (Not So) Hidden Dangers of Age Verification*, Electronic Frontier Foundation (Dec. 25, 2025), https://www.eff.org/deeplinks/2025/12/10-not-so-hidden-dangers-age-verification.
[14] CNIL, *Online Age Verification: Balancing Privacy and the Protection of Minors* (Sept. 22, 2022), https://www.cnil.fr/en/online-age-verification-balancing-privacy-and-protection-minors.
[15] This paper does not foreclose the possibility of privacy-preserving architectures using government identification, including zero-knowledge proofs built on top of digital credentials. Under such a system, a user's government ID might be presented once to a trusted credential issuer; thereafter, the user presents only a cryptographic proof of age eligibility to any given platform, which learns nothing about the user's identity, receives no document to retain, and cannot link that transaction to any other. The structural privacy risks that concern this paper would be sharply reduced under such a regime.

not be used or disclosed for any purpose other than age verification; third-party vendors must provide written assurances of confidentiality, security, and deletion, with no secondary use permitted; verification data must be promptly deleted after the age determination is made; clear notice must be provided to parents and users in the platform's privacy policy; reasonable security safeguards must be in place; and the method must produce reasonably accurate results across the full user population.

That last condition, accuracy across the full user population, is less straightforward than it sounds and deserves more regulatory attention than it has received. Accuracy rates for facial estimation vary substantially across demographic groups, and a method that performs well on average while systematically failing on specific populations does not meet the requirement of reasonable accuracy across the full user population in any meaningful sense.

Acceptable Tier 1 Methods

For Tier 1 high-risk features, where access must be meaningfully restricted to verified adults and where likelihood of harm to a child is high, the acceptable methods are those that provide a reliable age signal while minimizing the exposure of the user's underlying identity.

Zero-knowledge proofs and double-blind architectures represent the current privacy gold standard. In a zero-knowledge system, the age verification provider confirms only that the user meets a specified age threshold without learning the user's identity, and the service provider learns only the result of that confirmation without learning the user's identity from the verification provider. Neither party learns who the user is; they learn only that the user is age-eligible.[16] This architecture is technically demanding to implement correctly, and it is not yet widely deployed at scale, but it is available and becoming more accessible.[17] Age tokens and reusable credentials, where a user undergoes an initial verification process and receives a credential that can be presented to multiple services without re-exposing the underlying identity information, offer similar privacy properties when paired with appropriate user-binding mechanisms such as passkeys or biometrics.[18]

Device-level operating system signals offer a different architecture with similarly strong privacy properties. Rather than requiring the users to prove their age to a third-party verification provider, device-based systems allow the operating system itself to signal age eligibility to requesting services. The service learns only that a user's device has associated an age-eligibility flag; it does not learn the user's identity, and the user does not need to trust a third-party provider with personal data. This architecture is not yet a universal standard, but the

---

[16] Future of Privacy Forum, *Unpacking Age Assurance: Technologies and Tradeoffs* (v. 2.0) [hereinafter *FPF Age Assurance Infographic*] https://fpf.org/wp-content/uploads/2026/02/FPF-Age-Assurance-v2.0.pdf.

[17] Najmeh Tima, *Age Assurance Technologies, Emerging Standards, and Risk Management*, Captain Compliance (Mar. 6, 2026), https://captaincompliance.com/education/age-assurance-technologies-emerging-standards-and-risk-management/; *see also KGI Technical Assessment*, at 75–76 (discussing ZKP-based digital ID systems and noting Google's deployment of a ZKP-based age assurance system on Android).

[18] *FPF Age Assurance Infographic.*

privacy properties of device-based systems are strong precisely because the data does not travel through a third-party intermediary.

Methods that are not acceptable alone for Tier 1 include self-declaration, which is trivially circumvented by any motivated user; facial estimation without a secondary signal, which is both inaccurate near the age threshold and highly vulnerable to injection attacks using deepfake technology; and behavioral inference alone, which cannot provide any signal on first contact with a new user. These methods may play supporting roles in multi-signal systems, but none provides a sufficient basis for Tier 1 access control on its own.

Acceptable Tier 2 Methods

Tier 2 medium-risk features involve ongoing service relationships and lower circumvention incentives, which significantly expands the range of acceptable methods.

Behavioral and account-based inference is viable for Tier 2 when a behavioral profile already exists. A user's history of activity, the nature of one's connections, one's interaction patterns, and similar signals can, in aggregate, provide a reasonable basis for age estimation in an ongoing relationship. This method is not suitable as a first-contact mechanism, and it cannot substitute for upfront verification at Tier 1, but it is a legitimate tool for the kinds of feature-level access decisions that Tier 2 requires.

Parental vouching with a verified parent account is practical for Tier 2 contexts, particularly in gaming and social environments where parental involvement in account setup is already normalized.[19] The caveat is that current systems have difficulty verifying the parental relationship as distinct from merely verifying that the vouching adult is over 18 — a limitation that should be disclosed transparently and addressed as the technology matures.[20]

Self-declaration paired with a secondary behavioral signal is acceptable at Tier 2, with the understanding that self-declaration alone is insufficient and that the secondary signal does meaningful work.

AI facial estimation with disclosed demographic accuracy limits and a mandatory appeal pathway is acceptable for Tier 2 given its broad availability and its significantly better completion rates than government ID-based methods. However, it must be deployed with full transparency about its known inaccuracies near the age threshold and across demographic groups, and it

---

[19] *FPF Age Assurance Infographic* (illustrating parental vouching in an online gaming context: where a 16-year-old user lacks a government-issued ID, the system offers parental vouching as a fallback — an adult who has been independently verified as over 18 asserts that the minor is 16, and the resulting age credential is cryptographically bound to the minor's device passkey so that the age signal is only released upon entry of the correct PIN, pattern, or local biometric).

[20] *See* Future of Privacy Forum, *The State of Play: Verifiable Parental Consent and COPPA*, at 32 (Nov. 2021) (summarizing stakeholder critiques and noting that parents themselves described current VPC mechanisms as "privacy theater" susceptible to circumvention by children locating a parent's wallet or entering false information) https://fpf.org/wp-content/uploads/2021/11/FPF-The-State-of-Play-Verifiable-Parental-Consent-and-COPPA.pdf.

must never be the only available method. Research documents that facial age estimation systematically underperforms for transgender and nonbinary users, darker-skinned users, and users with disabilities.[21] No Tier 2 deployment of facial estimation should be permitted unless an alternative pathway is available to any user for whom the facial estimation result is inaccurate, contested, or simply unavailable.

VI. The Knowledge Asymmetry Rule

The debate over age assurance has proceeded largely as though the central challenge is compelling platforms to acquire information they do not have. It is not. The central challenge is compelling platforms to *act on* information they already use.

The business model of most major consumer platforms depends on knowing, with considerable granularity, who their users are. Platforms infer age from behavioral signals; device characteristics; account activity; and purchasing patterns, and they then use those inferences to classify users into audience segments that are then sold to advertisers. A platform that serves a user child-directed advertising, recommends age-appropriate content, or routes a user into a "family-friendly" content category has, by definition, already made a determination about that user's likely age. It has acted on an age signal. The signal may be probabilistic rather than certain, and it may have been derived through inference rather than direct collection, but it is an age signal, and the platform's commercial operations depend on it.

The knowledge asymmetry problem arises when that same platform, facing regulatory obligations that attach to knowledge of a user's age, disclaims the very knowledge its advertising systems depend on. This is not a theoretical concern. It is the operating posture of much of the platform industry.[22] Companies have successfully argued in regulatory and legislative contexts that they lack sufficient certainty about user ages to trigger heightened child safety obligations, while simultaneously maintaining the targeting infrastructure that requires exactly that certainty to function. The law has largely permitted this inconsistency. It should not.

A platform that classifies a user as a child, or as likely to be a child, for any commercial purpose, including but not limited to advertising targeting, content recommendation, or audience segmentation, is deemed to have actual knowledge of that user's age for purposes of all applicable child safety obligations. *The commercial use of an age signal is evidence of its possession.* A platform cannot have it both ways. Codifying this rule would require a direct, self-executing statutory rule that commercial use of age categorization constitutes knowledge for child-protective purposes.

This rule does not replace the need for age assurance methods of the kind described in Section V. A platform that does not use behavioral targeting, or that genuinely cannot infer user age from its existing data, still needs a mechanism for adjusting its default behavior toward users

[21] *New America Age Verification Report.*
[22] Federal Trade Commission, *A Look Behind the Screens: Examining the Data Practices of Social Media and Video Streaming Services* (Sept. 2024),https://www.ftc.gov/system/files/ftc_gov/pdf/Social-Media-6b-Report-9-11-2024.pdf.

who may be minors. But for the large majority of platforms whose commercial operations already depend on age inference, the knowledge asymmetry rule reframes the policy question entirely. The ask is not that platforms build something new. It is that they stop pretending they don't already have what they need.

VII. Vendor Transparency and Demographic Accuracy

The age assurance vendor market is largely opaque. Service providers select age verification providers based primarily on cost and compliance optics rather than independent assessments of accuracy, availability, or privacy protection. Users have no visibility into which vendors their platforms use, how those vendors perform across demographic groups, or what happens to the data they provide. Regulators have almost no publicly available information on which to base meaningful oversight. This is not a sustainable state of affairs for a market that is being asked to serve as critical child safety infrastructure.

This paper calls for mandatory public disclosure by age verification vendors on all four dimensions used in this framework's technical assessment[23]: baseline accuracy (including accuracy rates by age cohort, gender, race, disability status, and other relevant demographic categories); circumvention resistance (what attacks the system is and is not designed to resist, and what testing methodology was used to assess resistance); availability (what populations cannot use the system, and what the measured completion rate is across demographic groups); and privacy (what data is collected, how long it is retained, who has access to it, and what the architecture-level privacy properties are).

The accuracy disclosure requirement is particularly important because averages obscure the populations that matter most. A facial estimation system that achieves a mean absolute error of 2.7 years across a general population, itself a significant error rate when the relevant threshold is 18, may have error rates of 4 or 5 years for darker-skinned users or transgender users whose appearance the training data was not built to handle accurately.[24] A regulatory framework that demands only average accuracy will systematically under-protect the children in those communities while over-burdening the adults.

There is also a market structure concern that warrants attention. Because users who have already verified their age with a given provider do not want to verify again with a second provider, service providers have an incentive to choose widely used age verification vendors which creates pressure toward market concentration. Recent research found that the top five providers covered nearly 75 percent of the age verification market in two studied states, with a single vendor representing more than half of all sites.[25] Concentration in this market amplifies every privacy risk: A breach at a dominant age verification provider does not expose one platform's users but potentially *every user who has ever verified their age anywhere that*

[23] *KGI Technical Assessment*, at 24-27.
[24] *Id* at 78-82.
[25] *Id* at 95-96.

provider operates. This risk should be part of any serious regulatory assessment of the age assurance vendor ecosystem.

VIII. Conclusion

Age assurance is a tool with real costs — to privacy, to access, to speech, and above all, to the legitimate users who are most likely to be deterred by methods that were not designed with them in mind. These costs do not mean that age assurance is never appropriate. They mean that it is only appropriate in the right circumstances, applied through the right methods, with the right accountability structures in place.

The framework proposed in this paper is an attempt to specify those circumstances, methods, and structures in a way that is technically grounded, demographically honest, and constitutionally durable. It does not promise that age assurance will keep children away from content that harms them. It argues instead that age signals, properly used, can help platforms deliver experiences that are appropriate for the realistic range of users who will encounter them. This more modest, more achievable goal is actually more protective of children than the blocking approach that has dominated the policy conversation. Our four core asks are as follows:

First, adopt a three-tier feature classification system based on the technical distinction between features that must block minors and those that must provide safer defaults for users who may be minors. The method requirements for each tier should follow from what that distinction actually demands, not from what looks most rigorous in a legislative hearing.

Second, categorically exclude government ID as a required Tier 1 high-risk verification method. The privacy costs are too high, the structural exclusions too broad, and the completion rates too low. Privacy-preserving alternatives exist and are becoming more widely available. The regulatory framework should require the use of alternatives rather than defaulting to the most familiar and most harmful option of government ID checks.

Third, mandate public vendor transparency and demographic accuracy disclosure across all four assessment dimensions — accuracy, circumvention resistance, availability, and privacy — so that regulators, service providers, and users can make informed decisions about the systems that are being deployed in the name of child safety.

Fourth, codify the knowledge asymmetry rule: A platform that uses age signals to target users commercially for child-directed advertising or recommendations cannot simultaneously claim ignorance of those users' ages when facing regulatory obligations for child safety. The law should not permit companies to have it both ways.

This paper is a companion to Public Knowledge's broader child safety framework. The first paper made the case that safety by design is the most important lever available to policymakers. This paper fills the gap that the companion paper intentionally left open: when age assurance is part of the answer, here is how to do it right.