Prosocial Tech Design Regulation — *A Practical Guide*

By Lena Slachmuijlder

Council on Technology and Social Cohesion
Search for Common Ground
USC Marshall School of Business Neely Center for Ethical Leadership and Decision Making
Integrity Institute
Techsocietal
Prosocial Design Network
BUILD UP
goodbot

Foreword & Acknowledgements

Inspired by the Blueprint on Prosocial Tech Design Governance this Practical Guide offers guidance to regulators and civil society partners working towards tech platform governance. It captures insights from a diversity of technologists, peacebuilders, researchers and policy influencers associated with the Council on Tech and Social Cohesion and other peer organizations working towards safe and prosocial digital spaces.

Heartfelt thanks to the many people who contributed their expertise to strengthening the quality and relevance of this Practical Guide: Caleb Gichuhi (Build Up); Devika Malik; Guy Banim (Build Up); Habibou Bako (Search for Common Ground); Helena Puig Larrauri (Build Up); Ilamosi Ekenimoh (Integrity Institute); Julia Kamin (Prosocial Design Network); Lisa Schirch (University of Notre Dame); Marie-Eve Nadeau; Meghan Brown (Build Up); Peter Chapman (Knight-Georgetown Institute); Ravi Iyer (USC Neely Center); Renee Black (GoodBot); Samaya Anjum (Global Network Initiative); Sami Jaber, Scott Timcke (Research ICT Africa); Shruti Das (Columbia University); Sofia Bonilla (Integrity Institute); Spencer Gurley (Integrity Institute); Tope Ogundipe (TechSocietal).

For Citation

This publication should be cited as follows:
Slachmuijlder, Lena (2025). Prosocial Tech Design Governance: A Practical Guide. Brussels: Search for Common Ground.

Table of Contents

Executive Summary

Design shapes outcomes. Online harms don't just come from "bad content" or "bad people"—they flow from product design choices. Infinite scroll and autoplay drive overuse. Engagement-based ranking amplifies outrage, sensationalism, and polarization.[1] Open direct messages and frictionless tagging enable impersonation and abuse. Regulators, through this Practical Guide, will be able to implement clear, enforceable, content agnostic design governance.

Five Recommendations For Regulators:

1. **Stop manipulative, addictive design.** Ban the "dark patterns" that trick people into spending more time or sharing more data than they intend. Limit or allow users to switch off features like infinite scroll and autoplay, and set clear daily limits for high-volume actions. Require platforms to include time-management tools such as age-appropriate prompts to pause or take breaks, and give users the option to set and adjust their own daily or session time limits—all presented neutrally to help people, especially children and teens, make informed choices about how they spend time online.

2. **Make services safe by default for children.** When it is determined an account belongs to a child or teenager under 18, safer settings should automatically be turned on. This includes making profiles private by default, turning off autoplay and infinite scroll, limiting messages from strangers, and silencing notifications overnight. Do a child-impact assessment before any key design changes.

3. **Better recommenders, not just more engagement.** Require platforms to offer a user-selectable feed that prioritizes long-term value—credible information, bridging content (see definition below), and user preferences (e.g., "show more or less"), rather than pure engagement. Make this mode the default for children and teens, and ensure their choices are saved over time. Offer a non-profiling alternative feed. Platforms should clearly state what their feeds aim to optimize, test changes over several months, and publish simple, public summaries of the results.[2]

4. **Test before launch and be transparent.** Before any major design or ranking change, platforms must: (1) write a one-page plan in plain language explaining what's changing and define what "success" means; (2) test it on a small group first; (3) publish a short note summarizing what changed and what happened; and (4) allow an independent auditor to review the results.

---

[1] Stray, Jonathan, et al. *The Algorithmic Management of Polarization and Violence on Social Media.* Knight First Amendment Institute, August 2023.

[2] Cunningham, Tom, et al. "What We Know About Using Non-Engagement Signals in Content Ranking." arXiv, February 2024.

5

Measure success with user experience. Launch a multi-stakeholder initiative—convened by regulators with academics and civil society—to run rolling user panels and brief in-app pulse surveys tracking time well spent, credibility, exposure diversity, unwanted contact, and sleep disruption. Share privacy-safe findings regularly with platforms and tie them to recent design changes, using these insights to steer ongoing engagement toward greater health, safety, and trust among users.

Alongside mandates, regulators and civil society should advocate for voluntary uptake by digital platforms of evidence-based prosocial designs—norms reminders, rewarding quality posts, and accuracy prompts—that lift conversation quality without policing speech.[3]

Apply the same rules to new AI technologies. Chatbots, AI assistants, and generative AI tools shape what people see and believe just like feeds do. These systems should follow comparable duties, adapted to their risk level: safety-by-default (especially for minors), clear labeling and provenance for AI-generated media, proportionate friction on high-risk actions (e.g., mass AI messaging), pre-launch risk and impact assessments, public change-logs, independent audits, and privacy-safe researcher access.

Ensure regulations are rights-respecting. These measures are content-agnostic, complement illegal-content rules and digital-literacy efforts, and are consistent with UNESCO's Platform Governance Guidance, UNDP's Information Integrity Framework, UN's Convention on the Rights of the Child, the Abidjan Principles, the African Union's Policy on Child Online Safety, and the OHCHR's Online Platform Governance & Human Rights.[4,5,6,7,8]

Through this Practical Guide, regulators can ensure addictive mechanisms are addressed, safer defaults are implemented, recommendations are safer and users feel a sense of agency for their online experiences on platforms.

---

[3] Grüning, David, and Julia Kamin. Prosocial Design in Trust and Safety. June 2025. *T&S Past, Present, and Future* arXiv preprint.

[4] UNESCO. Guidelines for the Governance of Digital Platforms: Safeguarding Freedom of Expression and Access to Information through a Multi-stakeholder Approach. 2023.

[5] United Nations Development Programme. Strategic Guidance: Information Integrity: Forging a Pathway to Truth, Resilience and Trust. February 2022.

[6] "Déclaration d'Abidjan." REFRAM & RIARC, April 2024.

[7] African Union. Child Online Safety and Empowerment Policy. Adopted by the 44th Ordinary Session of the African Union Executive Council, February 2024, Addis Ababa, Ethiopia.

[8] United Nations Office of the High Commissioner for Human Rights. Online Platform: Governance & Human Rights. September 2025.

1. How To Use This Practical Guide

1.1 Purpose & Scope

This Practical Guide is for regulators, policymakers, and civil-society partners seeking to reduce systemic online harms by targeting design choices within the platform's architecture, rather than endlessly chasing harmful content. This is not a speech code, and it does not relax duties to remove illegal content.

Rather, we recognize that content moderation alone is not enough. Moderation acts after harm spreads. It fails at scale, is easy to game, and works unevenly across languages and contexts. Even where the language is the same, local context often influences the meaning and use of words. It is controversial and prone to over-reach: governments can pressure platforms to take down content, and platforms may over-remove lawful speech to avoid risk. Content moderation also backfires—inviting censorship, producing mistakes, and eroding trust in both regulators and platforms.[9] Most of all, it leaves intact the design choices (autoplay, infinite scroll, engagement-only ranking) that make harmful material travel fastest. Design governance fixes those mechanics so safety is built in by default while rights are preserved.

This Practical Guide argues that a more effective path is upstream design accountability—regulating the mechanics of feeds, defaults, prompts, and incentive structures that define user behavior on social media platforms.

1.2 Prosocial Tech Design Governance

Prosocial design governance aims to set clear, testable rules for the mechanics that shape people's online experience—feeds and ranking, defaults and prompts, sharing flows, notifications, contact or tagging, and AI-mediated interfaces (chatbots, assistants, generative tools). It is content-neutral and rights-respecting.

The goal is simple: tune product design so wellbeing, user control, and social cohesion are the default. The Blueprint for Prosocial Tech Design Governance frames three levers: (1) prosocial design standards, (2) transparency and independent verification, and (3) incentives and market shaping.[10] This Practical Guide focuses on the first two—where regulators have the most immediate influence—to make progress measurable, while encouraging voluntary uptake by platforms alongside enforceable duties.

---

[9] Iyer, Ravi. "Content Moderation Is a Dead End." The Psychology of Technology Institute Newsletter, October 2022, psychoftech.substack.com/p/content-moderation-is-a-dead-end.

[10] Schirch, Lisa. "Blueprint on Prosocial Tech Design Governance." Council on Technology and Social Cohesion with University of Notre Dame and Toda Peace Institute. May 2025.

1.3 Human and Child-Rights Grounding

This Practical Guide reflects the direction of the most recent global guidance on platform governance: move from policing speech to governing systems—**through transparency, accountability, and design-level duties.** UNESCO's Guidelines for the Governance of Digital Platforms call for human-rights–based, risk- and systems-oriented governance, with states being transparent and accountable about the requirements they place on platforms; they explicitly recognize children's special status and reference GC25.[11] In parallel, UNESCO's User Empowerment and MIL Action Plan stresses that the burden cannot sit on users alone; regulators should embed user autonomy and control through policy, plus require independent audits and other accountability measures.[12] That is exactly what this Practical Guide operationalizes (e.g., choice of modes, design change-logs, audits, and safe research access).

On information integrity, UNDP's strategic guidance treats the risks as structural and urges proportionate, evidence-based, rights-respecting responses that target those system drivers rather than policing opinions.[13] Our three levers—design standards, recommender governance, and transparency & audits—operationalize that advice into measurable duties regulators can deploy.

Many countries around the world are calling for platform governance to be in line with human and child rights. For instance, the European Union's Digital Services Act (DSA) mandates systemic risk assessment and design-led interventions to protect minors, while the UK's Age-Appropriate Design Code (AADC) sets standard requirements for privacy-by-default and safer user interfaces for children.[14] Australia's Safety-by-Design initiative focuses on embedding safeguards and transparency into digital product architecture and Indonesia's PP TUNAS Regulation introduce youth-protective defaults and recommender transparency duties.

Similarly, African continental policy now treats these harms as system problems. The AU's Child Online Safety & Empowerment Policy calls for national action plans, and ties platform duties to design-level safeguards and transparency—evidence that regulators are expected to act upstream.[15] Complementing this regional approach, Rwanda's Child Online Protection (COP) Policy provides an early national model for implementing safety- and privacy-by-design and default principles.[16]

---

[11] UNESCO. Guidelines for the Governance of Digital Platforms: Safeguarding Freedom of Expression and Access to Information through a Multi-stakeholder Approach. 2023.
[12] UNESCO. Towards User Empowerment: A Multi-stakeholder Action Plan for Integrating Media and Information Literacy on Digital Platforms: Companion Document. 2025.
[13] United Nations Development Programme. Strategic Guidance: Information Integrity: Forging a Pathway to Truth, Resilience and Trust. 2022.
[14] United Nations. How We Protect Children's Rights with the UN Convention on the Rights of the Child. 2023.
[15] African Union. Child Online Safety and Empowerment Policy. 2024.
[16] Republic of Rwanda, Ministry of ICT. Child Online Protection Policy. June 2019.

At the global level, UN CRC General Comment No. 25 directs states to require child-rights impact assessments, ensure proportionate regulation of digital design and data practices, and mandate business due diligence—recognizing that platform design and operation can cause or contribute to rights violations.[17]

The Abidjan Declaration and Protocol call on platforms to design for diversity and safety by default protections for minors.[18,19] They also require practical transparency, including disclosures and change-notices written in plain, clearly defined language—so regulators have a reference point when drafting laws and can assess compliance consistently. Additionally, the Declaration ensures researcher access to data and APIs with privacy safeguards, giving regulators a standing forum to oversee and test design changes.

Central to this Practical Guide is the integration of the 5Rights Foundation's Child Rights by Design principles and the Children's AI Design Code, which advocate embedding child rights into the architecture of digital services from the outset.[20,21] These child-rights and human-rights duties extend to AI features by design—covering labeling, provenance, safer defaults for minors, and proportionate, evidence-led mitigation for AI-specific risks (hallucination, impersonation, sycophancy, delusion, automated persuasion).

1.4 Why Platforms Change Their Design

Platforms change their designs for three main reasons: (1) to comply with regulation, (2) to avoid liability from lawsuits, or (3) to make voluntary changes—often pre-emptive moves shaped by the threat of enforcement. They also routinely adjust designs to maximize engagement, which in turn, makes their products more addictive or attention-capturing. For example, TikTok suspended and ultimately withdrew its "rewards for watching" feature across the EU only after the Commission opened proceedings under the DSA.[22] Similarly, many of the by-default privacy and safety changes documented under the UK's Age-Appropriate Design Code came only after the Code created a credible enforcement environment.[23]

---

[17] United Nations Committee on the Rights of the Child. General Comment No. 25 (2021) on Children's Rights in Relation to the Digital Environment. March 2021.
[18] "Déclaration d'Abidjan." REFRAM & RIARC, April 2024.
[19] African Union. "Africa Has Become the First Region in the World to Implement a Child Online Safety and Empowerment Policy." May 2024.
[20] 5Rights Foundation. "Child Rights by Design."
[21] 5Rights Foundation. Children & AI Design Code. 2025.
[22] "TikTok Commits to Permanently Withdraw TikTok Lite Rewards Programme from the EU to Comply with the Digital Services Act." Shaping Europe's Digital Future, European Commission, August 2024.
[23] Children and Screens: Institute of Digital Media and Child Development. "Landmark Report on the Impacts of the UK Age-Appropriate Design Code on Digital Platforms." March 2024.

The pattern is clear: left to their own incentives, profit will often outweigh safety. Platforms tend to act only when external pressure makes inaction costlier than reform. That is why regulators must establish frameworks that rebalance incentives. Clear, enforceable design duties make protecting users and fostering healthier online experiences the path of least resistance—and a business imperative rather than an afterthought.

Beware performative compliance. Some measures—like requiring "alternative feeds," one-page transparency plans, or additional testing—can be easily gamed. Platforms may make alternative feeds deliberately dull, phrase plans to sound positive while hiding trade-offs, or claim they already run tests without measuring real wellbeing impacts. For example, A/B experiments are often designed to optimize short-term engagement rather than long-term safety or user well-being.

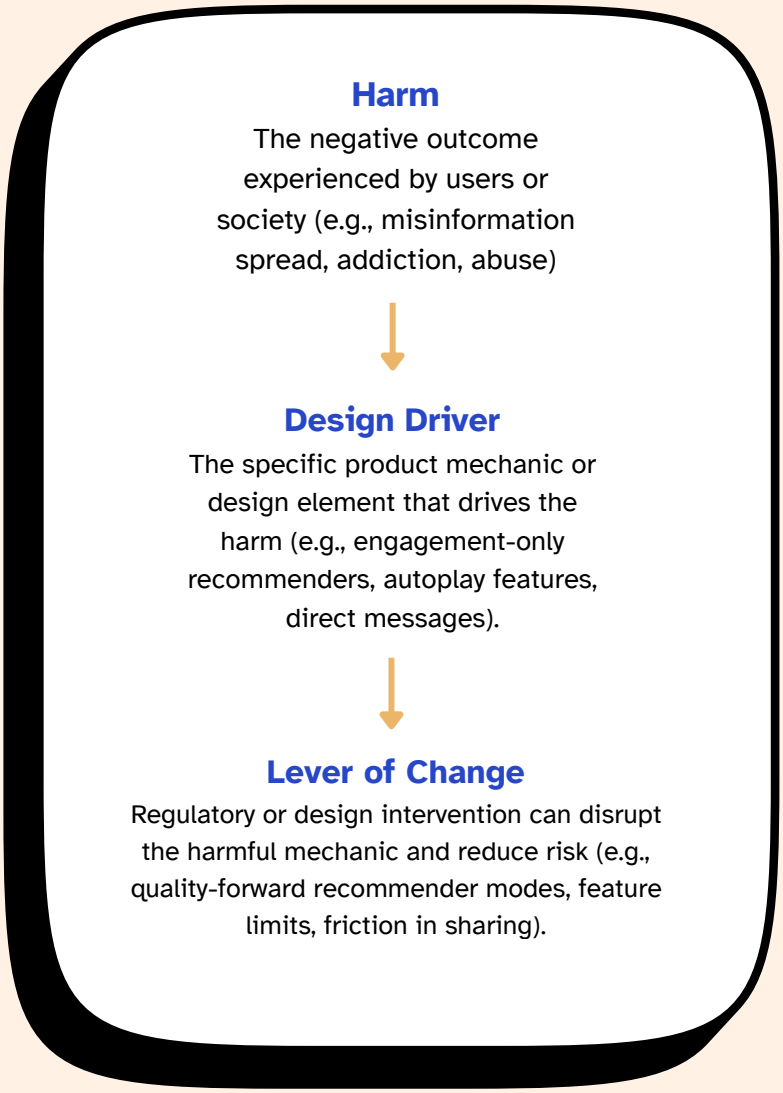
That's why regulators should focus on concrete, verifiable frontend or product-level rules, where compliance can be observed and measured, not just promised. Effective design governance requires clearly defined outcomes and robust oversight to ensure that platforms' actions produce meaningful improvements.

2. Connecting Harms to Design

Understanding why harm spreads, and how to prevent it, requires looking beyond individual content. The common thread across these findings is platform mechanics: features like engagement-only ranking, autoplay and infinite scroll, open contact and tagging amplify exposure while reducing user control. That's why this Practical Guide focuses on design levers, addressing the root causes of harm, rather than acting solely as a speech referee.

2.1 The "Harm → Design → Lever" Model

Platforms can harm users in primarily five ways: (1) excessive usage and addiction, (2) exposure to unwanted or harmful content, (3) harmful or unwanted contact, (4) privacy violations, and (5) manipulative or opaque recommender systems that amplify harmful content and shape long-term behaviors. The relationship between harm and design can be understood as a chain:



This model keeps the focus content-agnostic and viewpoint-neutral by concentrating on the mechanics of the product experience rather than the content or opinions themselves. It allows regulators and designers to pick the lightest yet effective levers to address specific harms.

2.1.1 Example No. 1 — Recommender Systems

What goes wrong? When feeds are tuned to maximize short-term engagement (clicks, likes, watch time), sensational and low-quality items rise fastest. That's the first-order problem: the model is targeting impulsive signals—a fact Mark Zuckerberg himself has highlighted—so one-click reshares and "next up" loops keep people in the same lane.[24] The second-order problem is incentives: product teams are rewarded for near-term engagement, so the system keeps learning toward outrage and novelty—even when users' longer-term preferences are for credible information, diverse viewpoints, and lower stress. The result: more sensational and polarizing content, plus more user regret.

How to fix it? Give people a clear "Better Feed" option that aims for long-term value—more credible information, bridging content, and less stress—rather than quick clicks.[25] Put likely minors in this mode by default; also offer more agency for users to select their algorithms, including a non-profiling or chronological option.

Then, consider changing what the system listens to: incorporate user feedback (e.g., surveys asking "what did you think of this content?"), apply quality metrics such as source reliability and original reporting, optimize for content that appeals to diverse audiences to reduce the visibility of polarizing material, and avoid chasing outrage.[26] Add small speed bumps on one-click reshares and display context or provenance panels so users see what they're sharing.

Finally, be clear and check the work: publish a short, plain-language note on what the feed is optimizing for, keep a public change-log, run longer tests (6–12 months), and allow independent audits—as recommended in KGI's Better Feeds report.[27]

---

[24] Zuckerberg, Mark. "A Blueprint for Content Governance and Enforcement." Facebook Notes, November 2018.
[25] Cunningham, Tom, et al. "What We Know About Using Non-Engagement Signals in Content Ranking." arXiv, February 2024.
[26] Slachmuijlder, Lena, and Sofia Bonilla. Prevention by Design: A Roadmap for Tackling Tech-Facilitated Gender-Based Violence at the Source. Search for Common Ground, Integrity Institute, and Council on Technology & Social Cohesion, March 2025.
[27] KGI (Knight Georgetown Institute). Better Feeds: Algorithms That Put People First. March 2025.

2.2.1 Example No. 2 — Addictive Overuse by Young People

What goes wrong? Compulsive use among teens is driven by interface mechanics that extend sessions by default: infinite scroll, “autoplay next,” and variable-reward loops that make stopping feel costly. Night-time push notifications can re-ignite sessions during vulnerable hours, contributing to sleep loss, next-day fatigue, and reduced attention in school or daily life. These effects are system-level, not tied to individual posts—so the most effective remedies focus on changing the mechanics, rather than policing the content itself.

How to fix it? Turn autoplay and infinite scroll off by default for likely minors, and apply a local night-curfew on notifications with a verified-adult override. Introduce brief late-night interstitials —such as a six-second pause with a simple breathing exercise—before users can continue scrolling or autoplaying.[28] These pauses are short enough to be unobtrusive but long enough to disrupt automatic, compulsive behavior, helping teens reflect on whether they want to continue and reducing negative impacts like sleep loss and next-day fatigue. Independent scans of the UK’s Age-Appropriate Design Code (AADC) period documented dozens of by-default changes of exactly this kind—autoplay reductions, night-notification curfews, and more private defaults—illustrating how clear design duties can shift population-level exposure.

[28] Danish Media Council for Children and Young People. Disrupting Social Media Habits: A Field Experiment with Young Danish Consumers. June 2025.

2.3 Mapping Harms → Design → Fix

Harm: User Experience	Design: Driver	Lever: What Can Regulators Do?
"I'm seeing fake or misleading news all over my feed."	Engagement-only ranking; 1-click reshares; speed-driven trending	Require a recommender system for long-term value (credibility & diversity over clicks); show provenance; publish objectives & change-logs
"I keep getting recommended harmful or dangerous content."	Outrage-tuned ranking; autoplay "related" loops	Shift ranking toward long-term value; set exposure-diversity guardrails; turn off autoplay for likely minors; publish testing results
"I can't stop scrolling or watching my feed."	Infinite scroll & autoplay maximize watch time	Default autoplay and infinite scroll off for likely minors; use pagination or a "next" button. Evidence shows brief pre-use tools help reduce compulsive engagement
"I'm getting notifications late and can't sleep."	Night-time push notifications	Set a default curfew (e.g., 00:00–06:00) with a verified-adult override
"Strangers keep messaging or tagging me, and some of it feels unsafe."	Open DMs; mass tagging; weak new-account throttles	Close DMs by default for minors; require consent-first tagging; apply graduated rate-limits for new accounts
"I'm being harassed or impersonated and there's no easy way to stop it."	Impersonation flows; no abuse triage	Guarantee a 24-hour impersonation redress path; provide a human-review channel; make evidence capture easy
"I feel tricked into sharing more data or agreeing to things I don't want."	Dark-patterned consent/settings	Ban deceptive patterns; run periodic design audits
"I only see what I agree with—or mostly sensational and divisionist content."	Signals optimized for clicks/anger	Require recommender system for long-term value; shift signals to diversity; publish change-logs
"I got AI replies that looked authoritative but were wrong."	Unchecked hallucination; no context.	Require AI-content labeling + provenance, sources/context panels, and pre-launch risk checks for high-impact AI features

3. How Regulation Can Shift Design

Independent evaluations show that clear, design-focused duties trigger concrete product changes at scale. In the UK, assessments of the Age-Appropriate Design Code (AADC) documented dozens of by-default changes across major services—autoplay reduced or off for minors, night-time notification curfews, more private settings, and stronger abuse filters.[29] A second synthesis tallied roughly 128 design and privacy changes between 2017–2024, with the steepest inflection coinciding with the AADC, EU, and UK systemic-risk regimes.[30]

There is also growing technical evidence that recommender systems can be redesigned to reduce harm without sacrificing service quality. The Better Feeds guidance synthesizes research and implementation experience to recommend: (1) user-selectable option optimized for long-term user value, (2) defaults that set minors into this long-term-value mode, and (3) long-horizon holdout experiments (12+ months) with public, auditable results.[31] **This incentivizes platforms to respond more to what users aspire to (often referred to as second-order preferences), rather than just what they click on (first-order preferences).**

Jurisdictions are increasingly integrating design governance into enduring oversight: systemic risk and design impact assessments before material changes, independent audits, public change-logs, and vetted researcher access (e.g., DSA Articles 34–40; UK Ofcom's evidence-led codes). These mechanisms ensure that claimed safety improvements can be tested and iterated, not just asserted. Based on such evidence, a key distinction arises between frontend product changes and backend algorithmic adjustments:

- **Frontend product changes** such as bans on autoplay, notification curfews, default privacy settings for minors, and break reminders—are implemented visibly in the user interface.[32] These changes are straightforward to verify and enforce because compliance is readily observable by users, auditors, or regulators. For example, a ban on autoplay can simply be tested by using the platform, and notification curfews can be checked by monitoring message delivery times.

- **Backend algorithmic adjustments**, which include recalibrating recommender systems to favor long-term user value over short-term engagement, introducing exposure diversity signals, or limiting amplification of sensational content, are less tangible. These changes occur "under the hood" and are inherently more complex to audit. Because these adjustments happen behind the scenes, verifying them typically requires access to the underlying algorithms and data—something only possible with greater transparency from platforms or controlled access for independent researchers.

---

[29] Information Commissioner's Office (ICO). Age-Appropriate Design: A Code of Practice for Online Services. United Kingdom, 2025.
[30] Wood, Steve. "New Laws and Regulations Around Child Safety and Privacy Raise Significant Questions." A&O Shearman Insights, October 2024.
[31] KGI (Knight Georgetown Institute). "Better Feeds: Algorithms That Put People First." March 2025.
[32] Kelly, Makena. "New Bill Would Ban Autoplay Videos and Endless Scrolling: Taking Aim at 'Features That Are Designed to Be Addictive.'" The Verge, July 2019.

Given these differences, regulatory strategies should prioritize enforceable frontend product-level mandates where possible, as they offer greater assurance of compliance and user benefit. At the same time, regulators should develop and require transparency frameworks and independent audit mechanisms to hold platforms accountable for backend algorithmic changes. This dual approach reflects emerging regulatory practices seen in the UK Age-Appropriate Design Code and the EU Digital Services Act, which combine visible interface restrictions with demands for algorithmic disclosures and audit access.

Bottom line: Clear, testable frontend design duties provide a critical baseline of safety and user protection, while backend recommender governance requires complementary transparency and oversight tools to ensure compliance and effectiveness. When regulators aim at design—defaults, recommender objectives and signals, velocity/virality controls, and auditable transparency—platforms make verifiable changes that improve safety at population scale. The record from the AADC, DSA enforcement, and Safety-by-Design shows this approach is feasible, repeatable, and rights-respecting.

3.1 Lessons from Other Industries

Societies have long turned to regulation to curb risks that markets alone fail to address. Seat belts in automobiles, health warnings on cigarettes, and drug safety testing were not voluntary innovations but mandated safeguards, introduced only after years of resistance and delay from industry. Before the introduction of seatbelts and cigarette regulations, societies faced alarming rates of preventable deaths and injuries, with countless lives lost to automobile accidents and smoking-related diseases that went unchecked due to the absence of mandatory safety measures and public health interventions.[33,34] These interventions demonstrate a consistent pattern: public safety rises to the forefront only when policy sets binding standards to address population-level risks.

Online platforms are now at a similar inflection point. Voluntary adjustments lag far behind the magnitude of harm facing vulnerable groups.[35] As with cars, tobacco, and pharmaceuticals, the most effective remedies come from upstream rules that embed safety, transparency, and accountability directly into core design.

Proactive, design-focused governance delivers measurable public benefits. Particularly relevant, would be the pharmaceuticals and financial service sectors, where systemic risk, transparency, and independent oversight are emphasized without dictating end products. By setting enforceable standards early, regulators can protect vulnerable populations, prevent avoidable harm, and secure broad social gains without compromising fundamental rights.

---

[33] Nolan, James. “Motor Vehicles—A Comparative Analysis of Seat Belt Legislation.” Cleveland-Marshall Law Review, vol. 14, no. 1, 1964, art. 12.

[34] Cunningham, Rob. “Tobacco Package Health Warnings: A Global Success Story.” Tobacco Control, vol. 31, no. 2, 2022, pp. 253–254.

[35] Frank Fagan, Systemic Social Media Regulation, 16 Duke Law & Technology Review 393-439 (2018).

3.2 Global Regulatory Postures

Regulators are shifting from policing speech to fixing design. Below are examples from different jurisdictions of how the regulators are looking upstream to design features, defaults, and recommender systems (algorithms).

3.2.1 EU — Digital Services Act (DSA)

- **Scope:** Very large online platforms/search. A systems approach: assess systemic risks (Art. 34) → mitigate them via design (Art. 35) → undergo annual independent audits (Art. 37) → give users recommender choice (Art. 38) → enable vetted researcher access (Art. 40). Minors get heightened protections (Art. 28).

- **Design Levers:** Risk-linked changes to ranking/virality/defaults; protections for minors (high safety, privacy-by-design, Art. 28); annual audits (Art. 37). Services must expose key recommender parameters to users (Art. 27).

- **Recommender Systems:** Offer at least one non-profiling option (Art. 38) and disclose the main parameters users can influence (Art. 27). This supports quality-forward/long-term-value modes, not just raw engagement.

- **Transparency & Access:** Run systemic risk assessments (Art. 34) and adopt reasonable, effective mitigations (Art. 35); publish risk summaries; undergo independent audits (Art. 37). Provide vetted researcher access under strict safeguards (Art. 40); the EU has adopted a delegated act detailing how data access should work.

- **What's Working?** The Commission's DSA action on TikTok Lite "rewards for watching" feature prompted TikTok to suspend it and, in August 2024, make binding commitments to remove the feature across the EU and not introduce a workaround.[36] This is an example of design-risk enforcement—targeting the mechanics rather than the content itself.

- **Replicable Design Levers:** Link risk findings to concrete design fixes (ranking signals, virality or velocity controls, defaults); center minors' protective defaults; offer a non-profiling feed people can pick; explain recommender parameters plainly; enable privacy-safe researcher access and independent audits.

[36] European Commission. "Commission Opens Proceedings Against TikTok under the DSA Regarding the Launch of TikTok Lite in France and Spain, and Communicates Its Intention to Suspend the Reward Programme in the EU." Press Release, April 2024, Brussels.

3.2.2 UK — Age-Appropriate Design Code Impact Assessment (AADC) & Online Safety Act (OSA)

- **Scope:** AADC has driven most substantial upstream design changes to protect children from online harms, under the oversight of the Information Commissioner's Office (ICO). Building on this foundation, the Online Safety Act (OSA), regulated by Ofcom, extends to a broader range of digital services and safety obligations.
- **Design Levers:** Privacy-by-default, safer contact and tagging, notification hygiene; system duties tied to user choice.
- **Recommender Systems:** Plain-language objectives; user-visible controls; mitigation testing.
- **Transparency & Access:** Product change-logs tied to codes and reporting guidelines.[37]
- **What's Working?** Independent scans show 128 by-default changes by platforms (autoplay down; night pings curbed; more private defaults).
- **Replicable Design Levers:** Make defaults do the work (best-interests/child-rights); tie risk assessments directly to design changes.

3.2.3 Australia — eSafety (Safety-by-Design)

- **Scope:** Applies to all online services accessed by users in Australia, including platforms where there is a significant risk of harm or abuse, including social networks, messaging services, and marketplaces.
- **Design Levers:** Special emphasis placed on protecting groups at higher risk from Technology-Facilitated Gender-Based Violence (TFGBV).
- **Recommender Systems:** Objective disclosure, user choice (chronological and interest-only), ongoing testing.
- **Transparency & Access:** Maintain a clear and up-to-date change-log documenting system modifications related to user safety.
- **What's Working?** Documented fixes after engagement; clearer recommender explanations in help centers.[38]
- **Replicable Design Levers:** Publish a template pack (SRA one-pager, audit prompts, change-log schema); use a transparent ladder that prompts design fixes before penalties.

---

[37] Ofcom. Online Safety Transparency Reporting: Final Transparency Guidance. United Kingdom, July 2025.
[38] eSafety Commissioner. Position Statement: Recommender Systems and Algorithms. Australia, December 2022.

3.2.4 USA — New York (SAFE for Kids Act and Child Data Protection Act) & Vermont (Act 63, 2025 AADC)

- **Scope:** Services likely to be used or accessed by minors (under 18) in New York and Vermont, including feeds, notifications, and child-data processing.
- **Design Levers:** Both frameworks restrict addictive or high-engagement feed mechanics without consent, setting overnight notification limits, and minimizing the collection and use of child data. Vermont's Act 63 adds specific age-assurance standards through Attorney General rulemaking.
- **Recommender Systems:** Require non-addictive or alternative curation for minors, transparent recommendation objectives, and easy-to-use controls. Engagement-only optimization goals for minors are discouraged.
- **Transparency & Access:** Both states mandate clear documentation and oversight—such as data protection or design impact assessments (DPIAs/DIAs), New York Attorney General may issue guidance and require periodic reporting.
- **What's Working?** Platforms have begun preparing feed controls and tightening notification defaults for minors, while early implementations in Vermont show progress toward more transparent consent flows and age-appropriate defaults.
- **Replicable Design Levers:** Treat feed and notification mechanics as design duties; integrate age-assurance and child-data minimization into product design; and tie default settings to child-rights and safety impact rationales.

3.2.5 USA — Minnesota (HF 4400 and Statute 325M.33)

- **Scope:** Social media platforms doing business in Minnesota or serving MN residents; transparency provisions effective July 1, 2025.
- **Design Levers:** Focus on design transparency around mechanics that drive harm: published interaction/velocity limits, quality & preference signals, notification practices (including night-hour stats), and product experiment summaries.
- **Recommender Systems:** Public explanation of how ranking systems use quality & expressed-preference signals and their relative weights (posted on the website).
- **Transparency & Access:** Statutory "algorithmic transparency" posting requirements (§ 325M.33); House Research summary notes AG enforcement and July 1, 2025 effective date.
- **What's Working?** New public dashboards/disclosures expected (e.g., percentile stats on user interactions; counts of night-time notifications; experiment descriptions).
- **Replicable Design Levers:** Mandate public, machine-readable posts on signals, weights, limits, notifications, experiments; tie enforcement to transparent, user-understandable recommender objectives.

3.2.6 Brazil — Digital Statute for Children and Adolescents (ECA Digital)

- **Scope:** Brazil's ECA Digital, enacted September 17, 2025, applies to any internet service, app, or technology product that is targeted at or likely to be accessed by minors under 18.
- **Design Levers:** Strict requirements for youth-protective product design and defaults: age verification, clear parental controls, restrictions on targeted advertising.
- **Recommender Systems:** Platforms must document the objectives of recommender algorithms, and limit youth profiling. Emphasis on exploring long-term value design choices, and minors' online data cannot be processed to intrude on their privacy.
- **Transparency & Access:** Maintaining public, semiannual reports on youth protection measures for providers with large numbers of underage users. Service providers must offer clear documentation and justification when removing content.
- **What's Working?** Brazil's pilot youth-protection efforts prompted platforms to begin implementing age assurance in advance of law's March 2026 effective date.
- **Replicable Design Levers:** Add recommender disclosure + youth defaults to rights baseline; phase-in researcher access.

3.2.7 Indonesia — PP TUNAS Regulation No. 17 of 2025

- **Scope:** PP TUNAS provides technical and operational rules for electronic system operators in Indonesia to protect children's personal data across digital platforms, complementing the broader Personal Data Protection (PDP) Law.
- **Design Levers:** Alignment with child-centred design standards; default protections for minors against manipulative design (early-stage).
- **Recommender Systems:** Encourages adoption of safer recommender modes such as chronological or non-profiling feeds for children, reducing exposure to harmful content and addictive design elements like autoplay and infinite scroll.
- **Transparency & Access:** Move toward design-risk framing; access mechanisms developing.
- **What's Working?** Policy statements signaling a pivot from pure moderation to design-first safeguards for minors.
- **Replicable Design Levers:** Codify minors' defaults (privacy high; pilot a chronological and non-profiling feed with public change-logs.

3.2.8 Korea — Protection Revision Act (Shutdown "Cinderella" Law)

- **Scope:** Applied to online games for minors under 16 in South Korea, the early law restricts access to digital gaming services during late-night hours to protect youth.[39]
- **Design Levers:** Banned minors from playing online games between 12:00AM and 6:00AM. In 2021, the Shutdown Law was abolished and replaced by a more flexible "selective game hours system" allowing parental control over gaming times.[40]
- **Recommender Systems:** While not explicitly legislated, platforms are encouraged to avoid addictive content loops and promote safer experiences for youth.
- **Transparency & Access:** Mandatory age verification and reporting by service providers.
- **What's Working?** Shutdown Law helped reduce late-night gaming but was criticized for pushing minors to circumvent rules; its abolition in 2021 aimed at empowering families with control rather than blanket restrictions.[41]
- **Replicable Design Levers:** Parental exemption and control mechanisms; age verification systems; algorithm moderation targeting addictive content

---

[39] Mielewczyk, Dominik Damian. "Korean Regulation of the Shutdown Law (셧다운제), and the Issue of Minors Using Electronic Games and Social Media." 2021.
[40] Bahk, Eun-ji. "Korea to Lift Game Curfew for Children." The Korea Times, August 2021.
[41] Hardawar, Devindra. "South Korea to End Its Controversial Gaming Curfew." Engadget, August 2021.

3.2.9 Building on the African Momentum

Why the context is ripe in Africa. Over the past decade, African countries have built serious scaffolding—cybercrime programs, data-protection and media regulators, alongside public investment in digital literacy and fact-checking. That progress now makes it feasible to prioritise prosocial design rules (feeds, defaults, virality controls) rather than relying solely only on content takedowns or user education.

The African Union's Continental Artificial Intelligence Strategy provides a foundational framework for mitigating online harms and promoting safe AI design across the continent.[42] It calls for comprehensive regulatory frameworks that prioritize risk mitigation and explicitly aligns AI policy with human rights principles. This strategy establishes legal protections and advocates for safety-by-design and transparency in AI systems.

Consider some regulatory highlights at the national level across the AU:

- **Kenya:** Computer Misuse and Cybercrimes Act (2018) criminalizes false publications and misinformation with penalties up to 10 years imprisonment or fines, addressing both intentional misinformation and its public harm.

- **Côte d'Ivoire:** Côte d'Ivoire has strengthened data protection with Law No. 2013-450, enforced by ARTCI, guaranteeing rights like data access, correction, and objection to processing. The country also combats fake news and misinformation through media literacy programs such as Désinfox Côte d'Ivoire, which train journalists and youth to verify information and reduce disinformation's impact on social cohesion.

- **Niger:** Niger's High Council for Communication (CSC) oversees broadcast and online news regulation to uphold media standards and pluralism. The country actively participates in ECOWAS-led efforts to harmonize cybercrime legislation and policies across West Africa. Furthermore, Niger has established a multi-stakeholder commission focused on information integrity, promoting collaboration among government, civil society, and private sector actors to combat misinformation and strengthen digital trust.

- **Senegal:** Strong CDP data-protection authority and long media-oversight tradition; misinformation debates paired with government-supported MIL and newsroom capacity-building.

[42] African Union. Continental Artificial Intelligence Strategy. August 2024.

- **Mali**: Mali's Autorité de Protection des Données à Caractère Personnel (APDP), established under Law No. 2013-015 and launched in 2016, serves as the national data protection authority. It actively balances security and rights through enforcement and regional collaboration, including participation in the Bamako Forum on Digital and Social Cohesion.

- **Nigeria:** Nigeria's Data Protection Act (NDPA 2023) builds on the foundational NDPR (2019) by establishing a comprehensive legal framework for data privacy, with stronger enforcement, enhanced user control, and sector-specific obligations.[43,44] The NDPA mandates registration, compliance audits, and data protection officers, with active enforcement already underway through the Nigeria Data Protection Commission (NDPC).

Voluntary regulator–platform commitments. The Abidjan Declaration commits platforms and regulators to: strong protections for minors; recommenders that favor diverse sources; practical transparency and researcher access and regional cooperation channel to monitor follow-through.[45]

Consider the following multi-stakeholder initiatives:

- The Policy Framework for Information Integrity in West Africa and the Sahel, adopted after the Regional Conference on Information Integrity in West Africa and the Sahel, The framework directs regulators to "focus on the systemic conditions enabling harmful and deceptive content to thrive" and adopts an upstream, systems stance. It notes that past measures often "fail to address the systemic drivers of disinformation and hate speech." It calls for platform-facing transparency and accountability—including conducting regular, independent algorithmic risk assessments." [46]

- The Bamako Forum on Digital & Social Cohesion (since 2023) convenes Burkina Faso, Mali, and Niger with media, platforms, and civil society; its 2024 Declaration pivots upstream—child-safe defaults, limits on autoplay/infinite scroll, clearer recommender objectives—while sustaining digital literacy and fact-checking.[47,48]

---

[43] Federal Republic of Nigeria. Nigeria Data Protection Act. 2023.
[44] Federal Republic of Nigeria. Nigeria Data Protection Act: Implementation Framework. 2019.
[45] "Déclaration d'Abidjan." REFRAM & RIARC, April 2024.
[46] UNESCO. Regional Conference on Information Integrity in West Africa and the Sahel. September 2025.
[46] "The Bamako Forum on Digital and Social Cohesion." SFCG (Bamako Forum).
[48] Forum de Bamako. Conclusions of the Third Edition of the Bamako Forum on Digital and Social Cohesion: Synergy of Action for Responsible Digital Policies. November 2024.

- The National Coalition on Freedom of Expression and Content Moderation (FeCoMo) in Kenya harnesses government, civil society and academic partnerships to advance a healthy and safe digital sphere.

- In September 2025 members of the Bamako Forum and FeCoMo published the 'Nairobi Declaration' calling for platform accountability at the system design level.[49]

- The Techsocietal African Online Safety Community of Practice has steadily advanced safety-by-design approaches with its member events.[50]

- Emergent research on harms, division and polarization are being experimented with by Build Up aimed at measuring the 'polarization footprint' of social media platforms, as an avenue to understanding its cost to society.[51]

Building on this momentum, authorities and civil society can invite platforms to pilot voluntary prosocial designs (prompts, exposure-diversity controls, session-break prompts) across Africa, with civil society and researchers co-evaluating outcomes for digital platform users as per the proposed metrics in Section 4.4.

[49] FECOMO, Search for Common Ground, Catalyst Forum, Bamako Forum. "Nairobi Call for a Safer Digital Future in Africa." Adopted in Nairobi, September 2025.
[50] TechSocietal. Online Safety Forum. Digital Access, Accountable Platforms & Inclusive Regulation. October 2025.
[51] Puig Larrauri, Helena. "Societal Divides as a Taxable Negative Externality of Digital Platforms." Build Up, March 2023.

4. Definitions & Legal Clauses

This library offers a shared design lexicon and copy-ready clause excerpts from existing laws. These offer practical, clear, enforceable and content-neutral rules to inspire future regulations.

4.1 Definitions — Design Lexicon

AI-Mediated Interface — Any feature that generates or meaningfully rewrites content or interactions for the user (e.g., chatbot, auto-reply, summarizer).

Addictive Mechanics — Patterns that prolong use without explicit user intent (e.g., infinite scroll, autoplay next), especially for minors.

Backend — Algorithmic and system-level processes that occur “under the hood,” including recommender system calculations, ranking adjustments, weighting of signals, and content amplification decisions.

Child-Specific Impact Assessment (CSIA) — An assessment applying the “best interests of the child” standard to foreseeable effects on minors.

Change Log — A record of all notable updates, fixes, and improvements made to a project, product, or piece of software over time.

Dark Pattern — An interface that materially subverts user autonomy (deception, obstruction, coerced choice).

Default — The pre-set configuration shown when the user takes no action.

Frontend — Visible, user-facing interface elements that can be directly observed and interacted with, such as autoplay settings, notification controls, break reminders, and privacy defaults.

Generative AI Content (Synthetic Media) — Text, image, audio, or video produced by an automated model, not recorded from the real world.

Long-Term-Value (LTV) mode — A user-selectable feed setting that prioritizes credibility, quality, source diversity, and wellbeing over short-term engagement.

Material Design Change — A change to recommender objectives/signals/weights; core defaults (privacy/notifications/contact); virality or velocity controls; or related experiments.

Non-Profiling Mode — A feed option that does not rely on behavioral profiling.

Content Provenance — A method of documenting and verifying the origin, history, and modifications of a piece of digital content, such as an image, video, audio file, or document.

Rate Limit — A documented cap on high-velocity actions (posting, resharing, tagging, invitations).

Recommender System — Any automated ranking, personalization, or suggestion of content or contacts.

Systemic Risk or Design Impact Assessment (SRA/DIA) — A pre-launch analysis of foreseeable risks from a design change and the mitigations chosen.

Technology-Facilitated Gender-Based Violence (TFGBV) — Any act of harm (physical, psychological, social, or economic) that is committed, assisted, amplified, or aggravated through digital technologies or online platforms and is directed at people based on their gender.

4.2 Levers of Change Regulators are Already Using

EACH ITEM = WHAT IT SOLVES → SHORT VERBATIM EXCERPT → SOURCE LABEL

4.2.1 Design Standards & Defaults

Limit who can find your account by default.
"by default set a user's account… to be discoverable only by the user's invited contacts or with the user's consent"
Source: Minnesota HF 4400, Subd. 3 ("Default Privacy Settings")

Limit geolocation visibility by default.
"by default… limit the ability of other account holders to know or share a user's geolocation to the minimum amount necessary."
Source: Minnesota HF 4400, Subd. 3(b)

Overnight notification curfew for minors.
"shall not knowingly send a push notification… to a minor during night-time hours, unless the user has opted in with a specific start and end time"
Source: New York SAFE for Kids Act (S7694-A)

High-privacy settings for children by default.
"Settings must be 'high privacy' by default"
Source: UK Age-Appropriate Design Code (ICO), Standard 7

Let children say "don't show me this" (negative feedback).
"Enable children to provide negative feedback on recommended content"
Source: UK Online Safety Act – Children's Code of Practice (RS3)

No "addictive feeds" for minors without checks.
"It is unlawful to provide an addictive feed to a covered user unless (A) the user is not a minor; or (B) there is verifiable parental consent."
Source: New York SAFE for Kids Act (S7694-A)

Cap high-velocity behaviors (spam and brigading control).
"must set and enforce strict, reasonable daily limits on… posts or reposts… new accounts followed… tagging frequency… and other actions with potential to cause harm through overuse."
Source: Minnesota HF 4400, Subd. 2

EACH ITEM = WHAT IT SOLVES → SHORT VERBATIM EXCERPT → SOURCE LABEL

4.2.2 Recommender Governance — Better Feeds

Offer a non-profiling feed option.
"Recipients of the service shall have at least one option for recommender systems not based on profiling."
Source: EU Digital Services Act, Article 38

Optimize for quality & expressed preferences (not raw engagement).
"the algorithmic ranking system must optimize content… that (1) a varied set of account holders indicates is of high quality; and (2) complies with a user's expressed preferences."
Source: Minnesota HF 4400, Subd. 1

Reduce youth targeting through suggestions.
"Refrain from the use of targeted or suggested groups, accounts, users, services, posts, and products on the youth account."
Source: Wisconsin S385 on Youth Accounts

Reduce prominence and avoid recommending harmful content to kids.
"Ensure that content likely to be [primary priority harmful] is not recommended to children" and "is reduced in prominence on children's recommender feeds."
Source: UK Online Safety Act – Children's Code of Practice (RS1, RS2)

4.2.3 Transparency & Researcher Access

Public, plain-language algorithm posts (signals & weights).
"maintain a public page that explains… ranking signals and their relative importance"
Source: Minnesota Stat. 325M.33 (HF 4400)

Risk assessments and design mitigations.
"carry out risk assessments… and put in place reasonable, proportionate and effective mitigation measures"
Source: EU Digital Services Act, Articles 34–35

Conduct independent audits.
"[Very large platforms] shall be subject to independent audits."
Source: EU Digital Services Act, Article 37-39

Ensure vetted researcher access.
"provide vetted researchers access to data… for research that contributes to the detection, identification, and understanding of systemic risks"
Source: EU Digital Services Act, Article 40

4.3 Harm Reduction via Design Regulation

What User Experiences (Harm)	Design Mechanic (Driver)	Remedy in Law or Code	What Does Success Looks Like?
"My feed is full of low-quality or misleading stuff."	Engagement-only ranking; one-click reshares; speed-based trending	EU DSA Art. 38 non-profiling option; MN HF 4400 Subd.1 optimize for quality & expressed preferences	↑ credible-source; ↑ exposure diversity; ↓ reshare velocity; ↓ regret-clicks
"I keep getting pulled back late at night."	Night-time push notifications	NY SAFE (S7694-A) overnight push limits for minors	↓ night notifications; ↑ reported sleep quality for minors
"I can't stop scrolling/watching."	Infinite scroll & autoplay	NY SAFE (S7694-A) (addictive-feed curb for minors); UK AADC high-privacy/child-safe defaults	↓ session length ↑ use of pagination or break prompts
"Strangers keep tagging or messaging me."	Open DMs; mass tagging; weak new-account throttles	MN HF 4400 Subd. 2 daily limits & stricter limits for new accounts	↓ unsolicited DMs ↓ first-week harassment from new accounts
"I'm seeing harmful content recommended to my kid."	Youth exposure via recommender	UK OSA children's code (RS1/RS2) do not recommend / reduce prominence; DSA Art. 38 non-profiling	↓ child exposure to flagged categories ↑ effective negative-feedback use
"I can't find/keep the settings I want."	Dark-patterned UX; low-privacy defaults	UK AADC "high privacy by default"; MN HF 4400 Subd.3 discoverability/geolocation limits by default	↑ settings stick rate; ↑ perceived control
"Spam/brigading overwhelms comments."	High-velocity posting, resharing, tagging	MN HF 4400 Subd. 2 daily rate limits	↓ bursts from new accounts ↓ coordinated harassment spikes
"No one can verify platform claims."	Opaque systems; no access	DSA Arts. 34–35, 39–40 risk assessments, audits, vetted researcher access	public change-logs; independent audit findings; published, replicable research

4.4 User Experiences as Success Metrics

The true test of design fixes is whether people's everyday online experiences actually improve, not just what the dashboards say. Major social platforms and regulators are already embracing this approach through regular user-experience surveys.[52]

- Snap Inc. runs an annual Digital Well-Being Index Study surveying teens and young adults across six countries (covering all their online activity, not just Snapchat) as a measure of Gen Z's psychological well-being.[53]
- UK regulator Ofcom conducts an "Internet Users' Experience of Harm Online" survey each year to quantify how often the public encounters content harms, privacy breaches, or security threats.[54]
- In Australia, the eSafety Commission's 2024 "Keeping Kids Safe Online" survey of 3,454 children found that 74% of kids aged 10–17 had seen harmful content and over half had experienced cyberbullying– underscoring the need for better design safeguards.[55]

Meanwhile, independent researchers have developed standardized indices. For example, the USC Neely Social Media Index uses rolling surveys to track whether users' positive vs. negative experiences are trending in the right direction on each platform.[56,57] Even the platforms themselves gather such data (e.g. Instagram's internal "Bad Experiences and Encounters Framework" user survey), though these findings rarely see the light of day.[58]

The solution is to make this feedback loop public, such as by **creating a regional, multi-stakeholder 'Digital Experience Observatory' to run an independent, rolling user panel and in-app pulse surveys.** Platforms participate, regulators convene, civil society co-governs, and researchers design the methods and curate public datasets.

---

[52] Ofcom: Appendix 1. 2024, Instagram. Bad Experiences and Encounters Framework (BEEF) Survey: Signals and Insights Platform. (July 2021).
[53] Snap Inc. Digital Well-Being Index – Year Three. February 2025.
[54] Ofcom. Technical Report: The Online Experiences Tracker (Wave 5). United Kingdom, January 2024.
[55] eSafety Commissioner. The Online Experiences of Children in Australia: Keeping Kids Safe Online Survey. Australia, 2024.
[56] Neely Center for Ethics & Technology. Neely Social Media Index. University of Southern California, 2025.
[57] Motyl, Matt, Jeff Allen, Jenn Louie, Spencer Gurley, and Sofia Bonilla. "Making Social Media Safer Requires Meaningful Transparency." Tech Policy Press, October 2024.
[58] Varanasi, Lakshmi. "Meta 'Misled' the Public Through a Campaign That Downplayed the Amount of Harmful Content on Instagram and Facebook, Court Documents Show." Business Insider, November 2023.

What to measure? Below are 10 examples of simple questions a user-centric survey might ask regularly – touching on everything from content quality to personal safety, to gauge if recent design changes are making a real difference:

- **Time Well Spent:** “After using [app] today, do you feel better or worse than prior?”
- **Content Credibility:** “How often do you see posts you consider well-sourced?”
- **Bridging Content Viewpoints:** “In the last week, did you see varying views on topics you follow?”
- **Recommendations Match Long-Term-Value:** “Does your feed reflect what you’ve said you want?”
- **Controls that Stick: “**When you change settings (notifications, who can DM you), do they stay as you set them?”
- **Sleep Disruption:** “Were you woken by late-night notifications this week?”
- **Unwanted Contact:** “How many times did someone you don’t know DM or tag you this week?”
- **Harassment or Impersonation:** “This week, did you experience or witness harassment or impersonation on [app]?” (Experienced / Witnessed / Both / Neither) → If experienced: “Was it easy to report and get help?”
- **AI Accuracy & Transparency (conditional):** “If you used an AI chatbot/search, did its answers provide sources you trust?”
- **Synthetic Media Clarity (conditional):** “Could you easily tell when an image, video, or text was AI-generated/edited?”

Regulators shouldn't have to guess if a new safety rule or design tweak is working – users will tell them, if asked. This kind of continuous, independent user-experience panel – paired with simple in-app survey pulses – is practical and within reach. This feedback loop keeps everyone honest: if the numbers aren't moving in the right direction, regulators and platforms will know it's time to adjust course.

User surveys can be complemented by other metrics and measurements that draw on recognized standards and evaluations. This includes compliance with recognized standards (such as IEEE2089: Standard for an Age Appropriate Digital Services Framework Based on the 5Rights Principles for Children), assessed through independent third-party accreditation audits.[59] Additionally, evaluations conducted by external experts—covering online interfaces, systems, settings, tools, functionalities, and reporting, feedback, and complaints mechanisms—provide another layer of accountability.

To ensure these user-experience and other metrics drive real accountability, an observatory could include regulators, academics, civil society representatives, and youth advisors to reflect diverse perspectives. Platforms would receive "safe harbor" protections when sharing data with the observatory, and accredited researchers could be granted privacy-preserving data access to dig deeper into the results.

[59] IEEE. IEEE 2089-2021: Standard for an Age-Appropriate Digital Services Framework Based on the 5Rights Principles for Children. 2021.

5. Conclusion

This Practical Guide starts from a simple truth: online harms are not inevitable—they are the predictable result of product choices. When regulators focus on design mechanics rather than policing opinions, safety becomes measurable, rights-respecting, and scalable. The evidence from child-safety regimes, DSA enforcement, and Safety-by-Design shows that clear, testable rules on feeds, defaults, notifications, tagging, and transparency trigger real product changes—especially for children and teens—without suppressing lawful speech. These changes reduce the spread of fake news and polarizing content, and reduce the incentives for people to post divisive and harmful content. The path forward is practical and immediate, requiring platforms to:

1. **Stop manipulative design and curb addictive mechanics.**

2. **Make services safe by default for minors.**

3. **Offer better feed options, which optimize for long-term value across credibility, bridging content, and user preferences.**

4. **Test material changes before launch, publish plain-language change-logs, and submit to independent audits and privacy-safe researcher access.**

5. **Measure success by user experiences, thus strengthening multi-stakeholder engagement in understanding the challenges and measuring progress together.**

Apply the same duties to AI assistants and generative features. Favor visible, front-end mandates (that anyone can verify) and pair them with strong transparency and audit obligations for back-end recommender governance, along with user-experience metrics. Many countries are ripe for an upstream, design-governance approach—anchored not only in African momentum but in a growing global consensus.

This approach builds on international frameworks and policies such as UNESCO's Guidelines for the Governance of Digital Platforms, the UN CRC General Comment No. 25, the UK's AADC, Australia's Safety-by-Design, and the Praia Policy Framework adopted at the September 2025 Regional Conference in Cape Verde conference. These policies and regulations call for multi-stakeholder cooperation and a focus on systemic, design-driven harms. Crucially, these frameworks anticipate and enable **meaningful collaboration with civil society and researchers,** who stand ready to co-evaluate outcomes so platforms can prove, not merely claim, that design changes improve everyday user experience.

If regulators set these duties now, platforms will tune towards wellbeing, agency, and social cohesion by default. Designing for better outcomes—rather than moderating after the fact—is how we protect rights, reduce harm at population scale, and rebuild trust in the digital public sphere.

6. Annex: Evidence of Digital Harms Across Africa

Across Africa, eight years of peer-reviewed studies, NGO reports, and law-enforcement assessments show platform harms are widespread and patterned. The biggest everyday problems: excessive or compulsive use, risks to children, tech-facilitated gender-based violence, and scams —often amplified by design choices like engagement-optimized feeds, infinite scroll or autoplay, open DMs, and the lack of built-in fact-checking or content verification tools.

- **Addiction:** A 2022 systematic review and meta-analysis of 22 African studies found a pooled internet-addiction prevalence of ~40% (higher in university cohorts), underscoring that heavy, hard-to-control use is common across the continent—not a "Western-only" issue.[60]

- **Harms to Women & Girls (TFGBV):** The African feminist study Alternate Realities, Alternate Internets documents routine online abuse; in its sample, 28% of women reported experiencing OGBV and 71% of reported incidents occurred on Facebook (with WhatsApp, Twitter/X, Instagram also prominent).[61] Plan International's global study (including African countries) found 58% of girls faced online harassment, often pushing them out of public conversation.[62]

- **Scams & Fraud:** INTERPOL identifies online scams as the number-one cyber threat reported by African member countries, highlighting phishing/social-engineering, fake investment schemes, and platform-mediated fraud that leverage social networks and messaging apps at scale.[63]

- **Hate Speech & Polarization:** Engagement-led ranking and fast reshares intensify inflammatory narratives and narrow people's exposure to diverse viewpoints. Spikes often coincide with elections, crises, or communal violence. Monitors in the Sahel have flagged surges of anti-minority content (including anti-Fulani rhetoric) during conflict peaks. A landmark Kenya suit by the family of Ethiopian professor Meareg Amare Abrha—alleging Facebook's failures around hate/incitement contributed to his 2021 killing—was allowed to proceed in 2025, underscoring how amplification failures can have offline consequences.[64]

---

[60] Endomba, Francky Teddy, et al. "Prevalence of Internet Addiction in Africa: A Systematic Review and Meta-Analysis." PLoS ONE, vol. 18, no. 1, 2023.
[61] Pollicy. Alternate Realities, Alternate Internets: African Feminist Research for a Feminist Internet. August 2020.
[62] Plan International and CNN's As Equals. "Hundreds of Girls Say They Face Online Harm at Least Once a Month." July 2024.
[63] INTERPOL. "INTERPOL Report Identifies Top Cyberthreats in Africa." October 2021.
[64] Aliyu, Saminu Mohammad, et al. "HERDPhobia: A Dataset for Hate Speech Against Fulani in Nigeria." arXiv, November 2022, arxiv.org/abs/2211.14623.

About the Council on Tech and Social Cohesion

Launched in 2023, the Council on Tech and Social Cohesion brings together technologists, peacebuilders, academics, and policy influencers around a defining question: *What if technology could foster trust and collaboration instead of driving polarization and violence?*

With more than 70 members across 22 countries, the Council's cross-sectoral and global network advocates for prosocial tech policies, measures technology's impact on social cohesion, and elevates evidence of products and practices that promote safe, healthy, and cohesive societies.

Focusing on design as the critical lever for change, the Council bridges research, policy, and innovation—demonstrating how technology can strengthen trust, safeguard human dignity, and become a force for peace and collaboration.