Régulation du Design Téchnologique Prosocial — un Guide Pratique

Par 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

Avant-propos

S'inspirant du Plan directeur pour la Gouvernance de la Conception technologique prosociale, ce guide pratique offre des orientations aux régulateurs et aux partenaires de la société civile qui œuvrent pour une gouvernance des plateformes technologiques. Il rassemble les points de vue de divers experts en technologies, acteurs de la consolidation de la paix, chercheurs et décideurs politiques associés au Conseil sur les Technologies et la Cohésion Sociale et à d'autres organisations homologues qui s'engagent pour des espaces numériques sûrs et prosociaux.

Un grand merci à toutes les personnes qui ont contribué par leur expertise à renforcer la qualité et la pertinence de ce guide pratique : 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 (Université de 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 (Université Columbia) ; Sofia Bonilla (Institut d'intégrité) ; Spencer Gurley (Institut d'intégrité) ; Tope Ogundipe (TechSocietal).

Pour citer ce rapport

Cette publication doit être citée comme suit :
Slachmuijlder, Lena (2025). Prosocial Tech Design Governance: A Practical Guide. Brussels: Search for Common Ground.

Table des matières

Résumé exécutif

Le design influence les résultats. Les préjudices en ligne ne proviennent pas uniquement de « contenus de mauvaise qualité » ou de « personnes mal intentionnées » ; ils découlent aussi des choix de la conception des produits. Le défilement infini et la lecture automatique favorisent la surutilisation. Le classement basé sur l'engagement amplifie l'indignation, le sensationnalisme et la polarisation.[1] L'ouverture des messages directs et la facilité d'utilisation des tags permettent l'usurpation d'identité et les abus. Grâce à ce guide pratique, les autorités de régulation pourront mettre en œuvre une gouvernance du design claire, applicable et indépendante du contenu.

Cinq recommandations à l'intention des organismes de réglementation:

1. **Mettez fin aux pratiques de conception manipulatrices et addictives.** Bannissez les « techniques de manipulation » qui incitent les utilisateurs à passer plus de temps en ligne ou à partager plus de données qu'ils ne le souhaitent. Limitez ou permettez aux utilisateurs de désactiver des fonctionnalités comme le défilement infini et la lecture automatique, et fixez des limites quotidiennes claires pour les actions fréquentes. Exigez des plateformes qu'elles intègrent des outils de gestion du temps, tels que des incitations adaptées à l'âge pour faire des pauses, et offrez-leur la possibilité de définir et d'ajuster leurs propres limites de temps quotidiennes ou par session – le tout présenté de manière neutre afin d'aider les utilisateurs, notamment les enfants et les adolescents, à faire des choix éclairés quant à leur utilisation d'Internet.

2. **Sécurisez les services par défaut pour les enfants.** Lorsqu'il est établi qu'un compte appartient à un enfant ou un adolescent de moins de 18 ans, des paramètres de sécurité renforcés doivent être activés automatiquement. Cela inclut la confidentialité par défaut des profils, la désactivation de la lecture automatique et du défilement infini, la limitation des messages provenant d'inconnus et la mise en sourdine des notifications pendant la nuit. Procédez à une évaluation d'impact sur les enfants avant toute modification importante de la conception.

3. **Des recommandations plus pertinentes, et pas seulement plus d'engagement.** Exiger des plateformes qu'elles proposent un flux personnalisable qui privilégie la valeur à long terme : informations crédibles, contenus de transition (voir définition ci-dessous) et préférences de l'utilisateur (par exemple, « afficher plus » ou « afficher moins »), plutôt que le simple engagement. Faire de ce mode le mode par défaut pour les enfants et les adolescents, et s'assurer que leurs choix sont enregistrés. Proposer un fil de contenu alternatif non profilant. Les plateformes doivent clairement indiquer les objectifs d'optimisation de leurs fil, tester les modifications pendant plusieurs mois et publier des synthèses simples et publiques des résultats.[2]

4. **Testez avant le lancement et faites preuve de transparence.** Avant toute modification majeure de conception ou de classement, les plateformes doivent : (1) rédiger un plan d'une page en langage clair expliquant les changements et définissant ce que signifie le « succès » ; (2) le tester d'abord auprès d'un petit groupe ; (3) publier une brève note résumant les changements et leurs résultats ; et (4) permettre à un auditeur indépendant d'examiner ces résultats.

---

[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

Mesurez le succès par les expériences des utilisateurs. Lancez une initiative multipartite – réunissant les autorités de réglementation, les universitaires et la société civile – afin de mettre en place des panels d'utilisateurs réguliers et de mener de brefs sondages d'opinion intégrés à l'application. Ces sondages permettront de suivre le temps passé, la crédibilité, la diversité des expériences, les contacts indésirables et les perturbations du sommeil. Partagez régulièrement les résultats, dans le respect de la vie privée, avec les plateformes et intégrez-les aux récentes modifications de conception. Utilisez ces informations pour orienter l'engagement continu vers une meilleure santé, sécurité et confiance des utilisateurs.

Parallèlement aux obligations légales, les organismes de réglementation et la société civile devraient plaider en faveur de l'adoption volontaire par les plateformes numériques de dispositifs prosociaux fondés sur des données probantes – rappels des normes, valorisation des publications de qualité et incitations à la précision – qui améliorent la qualité des échanges sans pour autant contrôler la parole.[3]

Il convient d'appliquer les mêmes règles aux nouvelles technologies d'IA. Les chatbots, les assistants vocaux et les outils d'IA générative, tout comme les flux d'information, façonnent ce que les gens voient et croient. Ces systèmes doivent respecter des obligations comparables, adaptées à leur niveau de risque : sécurité par défaut (notamment pour les mineurs), étiquetage clair et traçabilité des contenus générés par l'IA, mesures de restriction proportionnées pour les actions à haut risque (par exemple, l'envoi massif de messages par l'IA), évaluations des risques et des impacts avant le lancement, journaux de modifications publics, audits indépendants et accès sécurisé pour les chercheurs.

Veillez à ce que la réglementation respecte les droits. Ces mesures sont indépendantes du contenu, complètent les règles relatives aux contenus illégaux et les initiatives d'éducation au numérique, et sont conformes aux Orientations de l'UNESCO sur la Gouvernance des Plateformes, au Cadre d'Intégrité de l'Information du PNUD, à la Convention des Nations Unies relative aux droits de l'enfant, aux Principes d'Abidjan, à la Politique de l'Union Africaine sur la sécurité des enfants en ligne et à la Gouvernance des plateformes en ligne et aux droits humains du HCDH.[4,5,6,7,8]

Grâce à ce guide pratique, les organismes de réglementation peuvent s'assurer que les mécanismes de dépendance sont pris en compte, que des options par défaut plus sûres sont mises en œuvre, que les recommandations sont plus sûres et que les utilisateurs ont le sentiment de maîtriser leurs expériences en ligne sur les plateformes.

---

[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. Comment utiliser ce guide pratique

1.1 Objet et portée

Ce guide pratique s'adresse aux organismes de réglementation, aux décideurs politiques et aux partenaires de la société civile qui souhaitent réduire les préjudices systémiques en ligne en agissant sur les choix de conception au sein de l'architecture des plateformes, plutôt qu'en traquant sans cesse les contenus préjudiciables. Ce guide n'est pas d'un code de conduite relatif à la liberté d'expression et il ne dispense pas de l'obligation des plateformes de supprimer les contenus illégaux.

Nous reconnaissons que la modération de contenu à elle seule est insuffisante. Elle intervient après la propagation du contenu préjudiciable. Elle est inefficace à grande échelle, facilement manipulable et son efficacité varie selon les langues et les contextes. Même lorsque la langue est identique, le contexte local influence souvent le sens et l'usage des mots. La modération de contenu est controversée et sujette aux abus : les gouvernements peuvent faire pression sur les plateformes pour qu'elles retirent du contenu, et les plateformes peuvent supprimer excessivement des contenus légitimes pour éviter les risques. La modération de contenu a également des effets pervers : elle encourage la censure, engendre des erreurs et érode la confiance envers les organismes de réglementation et les plateformes.[9] Surtout, elle laisse intacts les choix de conception (lecture automatique, défilement infini, classement basé uniquement sur l'engagement) qui permettent aux contenus préjudiciables de se propager le plus rapidement possible. La gouvernance de la conception corrige ces mécanismes afin que la sécurité soit intégrée par défaut tout en préservant les droits.

Ce guide pratique soutient qu'une approche plus efficace consiste à responsabiliser les concepteurs en amont, en réglementant les mécanismes des flux, les paramètres par défaut, les invites et les structures d'incitation qui définissent le comportement des utilisateurs sur les plateformes de médias sociaux.

1.2 Gouvernance de la conception technologique prosociale

La gouvernance du design prosocial vise à définir des règles claires et vérifiables pour les mécanismes qui façonnent l'expérience en ligne des utilisateurs : flux et classement, paramètres par défaut et invites, parcours de partage, notifications, contact et étiquetage, et interfaces basées sur l'IA (chatbots, assistants, outils génératifs). Elle est neutre vis-à-vis du contenu et respectueuse des droits.

L'objectif est simple : concevoir des produits qui privilégient le bien-être, l'autonomie des utilisateurs et la cohésion sociale. Le Plan directeur pour une gouvernance de la conception de technologies prosociales s'appuie sur trois leviers : (1) des normes de conception prosociales, (2) la transparence et la vérification indépendante, et (3) les incitations et l'orientation du marché.[10] Ce guide pratique se concentre sur les deux premiers leviers – sur lesquels les organismes de réglementation ont l'influence la plus immédiate – afin de rendre les progrès mesurables, tout en encourageant l'adoption volontaire par les plateformes parallèlement à des obligations contraignantes.

---

[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 Fondements relatifs aux droits humains et aux droits de l'enfant

Ce guide pratique reflète les orientations des plus récentes recommandations internationales en matière de gouvernance des plateformes : **passer d'une surveillance des discours à une gouvernance des systèmes, fondée sur la transparence, la responsabilité et des obligations dès la conception**. Les Principes directeurs de l'UNESCO pour la Gouvernance des Plateformes Numériques préconisent une gouvernance fondée sur les droits humains, axée sur les risques et les systèmes, et exigent des États transparence et responsabilité quant aux exigences qu'ils imposent aux plateformes ; ils reconnaissent explicitement le statut particulier des enfants et font référence à la Convention générale de l'UNESCO n° 25.[11] Parallèlement, le Plan d'action de l'UNESCO pour l'autonomisation des utilisateurs et l'éducation aux médias et aux ressources éducatives (EMRI) souligne que cette responsabilité ne peut reposer uniquement sur les utilisateurs ; les autorités de régulation doivent intégrer l'autonomie et le contrôle des utilisateurs dans leurs politiques, et exiger des audits indépendants ainsi que d'autres mesures de responsabilisation.[12] C'est précisément ce que ce guide pratique met en œuvre (par exemple, le choix des modes d'utilisation, la traçabilité des modifications apportées à la conception, les audits et l'accès sécurisé à la recherche).

En matière d'intégrité de l'information, les orientations stratégiques du PNUD considèrent les risques comme structurels et préconisent des réponses proportionnées, fondées sur des données probantes et respectueuses des droits, qui ciblent ces facteurs systémiques plutôt que de contrôler les opinions.[13] Nos trois leviers — normes de conception, gouvernance des organismes de recommandation et transparence et audits — transforment ces recommandations en obligations mesurables que les organismes de réglementation peuvent mettre en œuvre.

De nombreux pays à travers le monde appellent à une gouvernance des plateformes conforme aux droits humains et aux droits de l'enfant. Par exemple, la réglementation européenne sur les services numériques (DSA) impose une évaluation systémique des risques et des interventions dès la conception pour protéger les mineurs, tandis que le code de conception adapté à l'âge (AADC) du Royaume-Uni définit des exigences standard en matière de protection de la vie privée par défaut et d'interfaces utilisateur plus sûres pour les enfants.[14] L'initiative australienne « Safety-by-Design » vise à intégrer des garanties et la transparence dans l'architecture des produits numériques, et le règlement indonésien PP TUNAS introduit des paramètres par défaut protégeant les jeunes et des obligations de transparence pour les systèmes de recommandation.

De même, les politiques continentales africaines considèrent désormais ces préjudices comme des problèmes systémiques. La politique de l'Union Africaine relative à la sécurité et à l'autonomisation des enfants en ligne préconise des plans d'action nationaux et lie les obligations des plateformes à des garanties intégrées dès la conception et à la transparence, preuve que les régulateurs sont censés agir en amont.[15] Complétant cette approche régionale, la politique Rwandaise de protection des enfants en ligne constitue un modèle national précurseur pour la mise en œuvre des principes de sécurité et de protection de la vie privée intégrés dès la conception et par défaut.[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.

Au niveau mondial, l'Observation générale n° 25 de la Convention relative aux droits de l'enfant des Nations Unies enjoint aux États d'exiger des évaluations d'impact sur les droits de l'enfant, d'assurer une réglementation proportionnée de la conception numérique et des pratiques en matière de données, et d'imposer une obligation de diligence raisonnable aux entreprises, reconnaissant que la conception et l'exploitation des plateformes peuvent entraîner ou contribuer à des violations des droits.[17]

La Déclaration et le Protocole d'Abidjan invitent les plateformes à concevoir des environnements inclusifs et sécurisés, notamment pour les mineurs.[18,19] Ils exigent également une transparence concrète, avec des informations et des notifications de modification rédigées dans un langage clair et précis, afin que les autorités de régulation disposent d'un point de référence lors de l'élaboration des lois et puissent évaluer la conformité de manière cohérente. De plus, la Déclaration garantit aux chercheurs l'accès aux données et aux API avec des garanties de confidentialité, offrant ainsi aux autorités de régulation un cadre permanent pour superviser et tester les modifications de conception.

Ce guide pratique intègre les principes de la Fondation 5Rights relatifs aux droits de l'enfant dès la conception et le Code de Conception de l'IA pour les enfants, qui préconisent d'intégrer les droits de l'enfant dans l'architecture des services numériques dès le départ.[20,21] Ces obligations en matière de droits de l'enfant et de droits de l'homme s'étendent aux fonctionnalités de l'IA dès la conception, notamment l'étiquetage, la provenance, des valeurs par défaut plus sûres pour les mineurs et une atténuation proportionnée et fondée sur des preuves des risques spécifiques à l'IA (hallucination, usurpation d'identité, flagornerie, délire, persuasion automatisée).

1.4 Pourquoi les plateformes modifient leur conception

Les plateformes modifient leur conception pour trois raisons principales : (1) se conformer à la réglementation, (2) éviter d'être tenues responsables en cas de poursuites judiciaires, ou (3) apporter des modifications volontaires – souvent des mesures préventives motivées par la menace de sanctions. Elles ajustent également régulièrement leur conception pour maximiser l'engagement, ce qui rend leurs produits plus addictifs ou captivants. Par exemple, TikTok a suspendu, puis retiré, sa fonctionnalité « récompenses pour le visionnage » dans toute l'UE seulement après que la Commission a ouvert une procédure au titre de la loi britannique sur les services numériques (DSA).[22] De même, bon nombre des modifications apportées par défaut aux paramètres de confidentialité et de sécurité, documentées par le code britannique de conception adapté à l'âge, ne sont intervenues qu'après la mise en place d'un cadre juridique crédible.[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.

Le constat est clair : livrés à leurs propres intérêts, les plateformes privilégient souvent le profit à la sécurité. Elles n'agissent généralement que lorsque des pressions extérieures rendent l'inaction plus coûteuse qu'une réforme. C'est pourquoi les autorités de régulation doivent mettre en place des cadres permettant de rééquilibrer les incitations. Des obligations de conception claires et applicables font de la protection des utilisateurs et de la promotion d'expériences en ligne plus saines la voie la plus facile à suivre – et un impératif commercial plutôt qu'une simple considération secondaire.

Attention à la conformité de façade. Certaines mesures, comme l'exigence de « flux alternatifs », de plans de transparence d'une page ou de tests supplémentaires, sont facilement manipulables. Les plateformes peuvent volontairement rendre les flux alternatifs peu attrayants, formuler leurs plans de manière à paraître positifs tout en dissimulant les compromis, ou prétendre mener déjà des tests sans mesurer les impacts réels sur le bien-être. Par exemple, les tests A/B sont souvent conçus pour optimiser l'engagement à court terme plutôt que la sécurité ou le bien-être des utilisateurs à long terme.

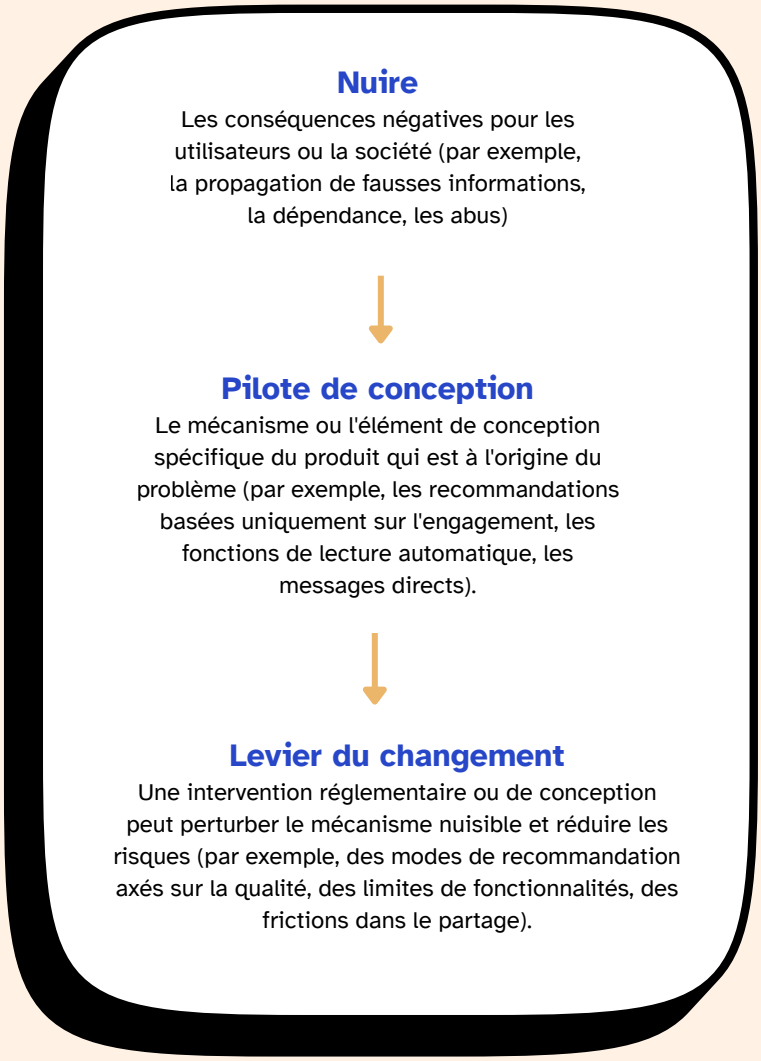
C'est pourquoi les organismes de réglementation devraient privilégier des règles concrètes et vérifiables au niveau de l'interface utilisateur ou du produit, où la conformité peut être observée et mesurée, et non simplement promise. Une gouvernance efficace de la conception exige des résultats clairement définis et une surveillance rigoureuse afin de garantir que les actions des plateformes produisent des améliorations significatives.

2. Lier les préjudices à la conception

Pour comprendre la propagation des contenus préjudiciables et comment la prévenir, il est nécessaire d'aller au-delà du simple contenu individuel. Le point commun de ces constats réside dans les mécanismes des plateformes : des fonctionnalités telles que le classement basé uniquement sur l'engagement, la lecture automatique et le défilement infini, l'ouverture des contacts et le système de tags amplifient la visibilité tout en réduisant le contrôle de l'utilisateur. C'est pourquoi ce guide pratique se concentre sur les leviers de conception, en s'attaquant aux causes profondes des préjudices, plutôt que de se contenter d'agir comme un arbitre du discours.

2.1 Le modèle « Dommage → Conception → Levier »

Les plateformes peuvent nuire aux utilisateurs principalement de cinq manières : (1) usage excessif et dépendance, (2) exposition à des contenus indésirables ou nuisibles, (3) contacts nuisibles ou non désirés, (4) atteintes à la vie privée et (5) systèmes de recommandation manipulateurs ou opaques qui amplifient les contenus nuisibles et influencent les comportements à long terme. La relation entre préjudice et conception peut être appréhendée comme une chaîne :



Ce modèle privilégie une approche indépendante du contenu et neutre vis-à-vis des points de vue, en se concentrant sur les mécanismes de l'expérience utilisateur plutôt que sur le contenu ou les opinions elles-mêmes. Il permet aux organismes de réglementation et aux concepteurs de choisir les leviers les plus simples et efficaces pour remédier à des problèmes spécifiques.

2.1.1 Exemple n° 1 — Systèmes de recommandation

Qu'est-ce qui cloche? Lorsque les flux sont optimisés pour maximiser l'engagement à court terme (clics, mentions « J'aime », temps de visionnage), les contenus sensationnalistes et de piètre qualité émergent le plus rapidement. C'est le problème de premier ordre : le modèle cible les réactions impulsives – un fait que Mark Zuckerberg lui-même a souligné – et les partages instantanés et les boucles « à suivre » maintiennent les utilisateurs dans la même tendance.[24] Le problème de second ordre réside dans les incitations : les équipes produit sont récompensées pour l'engagement immédiat, si bien que le système continue d'apprendre à privilégier l'indignation et la nouveauté, même si les préférences à long terme des utilisateurs vont vers des informations crédibles, des points de vue diversifiés et un environnement moins stressant. Résultat : davantage de contenus sensationnalistes et clivants, et davantage de regrets de la part des utilisateurs.

Comment y remédier? Proposer aux utilisateurs une option « Flux amélioré » claire, axée sur la valeur à long terme (informations plus crédibles, contenus complémentaires et réduction du stress) plutôt que sur des clics rapides.[25] Activer ce mode par défaut pour les mineurs potentiels et leur donner davantage de latitude pour choisir leurs algorithmes, notamment une option sans profilage ou chronologique.

Ensuite, envisagez de modifier les critères d'évaluation du système : intégrez les commentaires des utilisateurs (par exemple, des sondages du type « Qu'avez-vous pensé de ce contenu ? »), appliquez des indicateurs de qualité tels que la fiabilité des sources et l'originalité des reportages, optimisez le contenu pour qu'il s'adresse à un public diversifié afin de réduire la visibilité des contenus polémiques, et évitez de céder à la polémique.[26] Ajoutez de légères limitations au partage en un clic et affichez des panneaux de contexte ou de provenance pour que les utilisateurs voient ce qu'ils partagent.

Enfin, soyez clair et vérifiez votre travail : publiez une brève note en langage clair sur les objectifs d'optimisation du flux, tenez un journal des modifications public, effectuez des tests plus longs (6 à 12 mois) et autorisez les audits indépendants, comme le recommande le rapport « Better Feeds » de KGI.[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 Exemple n° 2 — Consommation excessive et addictive chez les jeunes

Qu'est-ce qui cloche? L'usage compulsif chez les adolescents est alimenté par des mécanismes d'interface qui prolongent automatiquement les sessions : défilement infini, lecture automatique et boucles de récompense variables qui rendent l'arrêt difficile. Les notifications nocturnes peuvent relancer les sessions pendant les heures les plus vulnérables, contribuant ainsi à des troubles du sommeil, à la fatigue du lendemain et à une baisse de l'attention à l'école ou dans la vie quotidienne. Ces effets sont systémiques et non liés à des publications individuelles ; les solutions les plus efficaces consistent donc à modifier ces mécanismes plutôt qu'à contrôler le contenu lui-même.

Comment y remédier? Désactivez par défaut la lecture automatique et le défilement infini pour les mineurs potentiels et instaurez un couvre-feu nocturne local pour les notifications, avec autorisation d'un adulte vérifié. Introduisez de brèves pauses en fin de soirée – comme une pause de six secondes avec un simple exercice de respiration – avant que les utilisateurs puissent reprendre le défilement ou la lecture automatique.[28] Ces pauses sont suffisamment courtes pour être discrètes, mais suffisamment longues pour interrompre les comportements automatiques et compulsifs, aidant ainsi les adolescents à réfléchir à la suite et réduisant les effets négatifs tels que le manque de sommeil et la fatigue du lendemain. Des analyses indépendantes de la période du Code de conception adapté à l'âge (AADC) au Royaume-Uni ont recensé des dizaines de modifications par défaut de ce type – réduction de la lecture automatique, couvre-feu nocturne pour les notifications et autres paramètres par défaut plus respectueux de la vie privée – illustrant comment des obligations de conception claires peuvent influencer l'exposition à l'échelle de la population.

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

2.3 Cartographie des préjudices → Conception → Correction

Préjudice : Expérience utilisateur	Conception : Moteur	Levier : Que peuvent faire les régulateurs ?
« Je vois de fausses informations ou des nouvelles trompeuses partout dans mon fil. »	Classement basé uniquement sur l'engagement ; partage en 1 clic ; tendances guidées par la vitesse	Un système de recommandation est nécessaire pour une valeur à long terme (crédibilité et diversité plutôt que clics) ; Afficher la provenance ; publier les objectifs et les journaux de modifications
« Je continue de recevoir du contenu nuisible ou dangereux recommandé. »	Outrage-tuned ranking; autoplay "related" loops	Prioriser le classement en fonction de la valeur à long terme ; définir des garde-fous de diversification de l'exposition ; désactiver la lecture automatique pour les mineurs potentiels ; publier les résultats des tests
« Je n'arrive pas à arrêter de faire défiler ou regarder mon fil. »	Défilement infini & lecture automatique maximisent le temps de visionnage	Désactivez par défaut la lecture automatique et le défilement infini pour les mineurs potentiels ; utilisez la pagination ou un bouton « Suivant ». Il a été démontré que des outils de pré-utilisation succincts contribuent à réduire l'engagement compulsif.
« Je reçois des notifications tard et je ne peux pas dormir. »	Notifications push nocturnes	Fixer un couvre-feu par défaut (ex. 00:00–06:00) avec dérogation pour adulte vérifié
« Des inconnus m'envoient des messages ou me taguent, et certains comportements semblent dangereux. »	Messages privés ouverts ; tags massifs ; limitations faibles pour les nouveaux comptes	Fermer les messages privés par défaut pour les mineurs ; exiger le consentement pour le tagging ; appliquer des limitations progressives pour les nouveaux comptes
« Je suis harcelé ou usurpé et il n'y a pas de moyen facile de l'arrêter. »	Flows d'usurpation ; absence de triage des abus	Garantir un recours en 24 heures pour l'usurpation ; fournir un canal de révision humaine ; faciliter la capture de preuves
« Je me sens trompé pour partager plus de données ou accepter des choses que je ne veux pas. »	Consentement, settings trompeurs (dark patterns)	Interdire les pratiques trompeuses ; effectuer des audits de conception périodiques
« Je ne vois que ce avec quoi je suis d'accord — ou du contenu surtout sensationnaliste et divisionniste. »	Signaux optimisés pour les clics, la colère	Exiger un système de recommandation basé sur la valeur à long terme ; réorienter les signaux vers la diversité ; publier les journaux de modifications
« J'ai reçu des réponses d'IA qui semblaient fiables mais étaient fausses. »	Hallucinations non vérifiées ; absence de contexte	Exiger l'étiquetage du contenu IA + provenance, panneaux de sources/contexte, et vérifications de risque avant lancement pour les fonctionnalités IA à fort impact

3. Comment la réglementation peut influencer la conception

Des évaluations indépendantes montrent que des obligations claires et axées sur la conception entraînent des modifications concrètes et à grande échelle des produits. Au Royaume-Uni, les évaluations du Code de conception adapté à l'âge (AADC) ont recensé des dizaines de modifications par défaut dans les principaux services : réduction ou désactivation de la lecture automatique pour les mineurs, couvre-feu nocturne pour les notifications, paramètres de confidentialité renforcés et filtres anti-abus plus performants.[29] Une seconde synthèse a dénombré environ 128 modifications de conception et de confidentialité entre 2017 et 2024, la plus forte augmentation coïncidant avec la mise en place de l'AADC, des régimes de risque systémique de l'UE et du Royaume-Uni.[30]

Il existe également de plus en plus de preuves techniques que les systèmes de recommandation peuvent être repensés pour réduire les risques sans compromettre la qualité du service. Le guide « Better Feeds » synthétise la recherche et l'expérience de mise en œuvre pour recommander : (1) une option sélectionnable par l'utilisateur, optimisée pour une valeur ajoutée à long terme ; (2) des paramètres par défaut qui activent ce mode de valeur à long terme pour les mineurs ; et (3) des expérimentations à long terme (plus de 12 mois) avec des résultats publics et vérifiables.[31] **Cela incite les plateformes à mieux répondre aux aspirations des utilisateurs (souvent appelées préférences de second ordre) plutôt qu'à leurs seuls clics (préférences de premier ordre).**

Les autorités intègrent de plus en plus la gouvernance de la conception dans un contrôle permanent : évaluations systémiques des risques et de l'impact de la conception avant toute modification importante, audits indépendants, journaux de modifications publics et accès contrôlé des chercheurs (par exemple, les articles 34 à 40 de la DSA ; les codes de conduite fondés sur des preuves de l'Ofcom au Royaume-Uni). Ces mécanismes garantissent que les améliorations de sécurité annoncées peuvent être testées et itérées, et non pas seulement affirmées. Sur la base de ces preuves, une distinction essentielle apparaît entre les modifications apportées au produit (interface utilisateur) et les ajustements algorithmiques (interface utilisateur).

- **Les modifications apportées à l'interface utilisateur,** telles que l'interdiction de la lecture automatique, les restrictions de notification, les paramètres de confidentialité par défaut pour les mineurs et les rappels de pause, sont intégrées de manière visible.[32] Ces modifications sont faciles à vérifier et à appliquer, car leur conformité est aisément observable par les utilisateurs, les auditeurs ou les autorités de réglementation. Par exemple, l'interdiction de la lecture automatique peut être testée simplement en utilisant la plateforme, et les restrictions de notification peuvent être vérifiées en surveillant les délais de livraison des messages.

---

[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.

- **Les ajustements algorithmiques en arrière-plan,** tels que le recalibrage des systèmes de recommandation pour privilégier la valeur à long terme pour l'utilisateur plutôt que l'engagement à court terme, l'introduction de signaux de diversification de l'exposition ou la limitation de la diffusion de contenus sensationnalistes, sont moins tangibles. Ces modifications s'opèrent « en coulisses » et sont intrinsèquement plus complexes à auditer. Comme ces ajustements se font en coulisses, leur vérification nécessite généralement un accès aux algorithmes et aux données sous-jacents — ce qui n'est possible qu'avec une plus grande transparence de la part des plateformes ou un accès contrôlé pour les chercheurs indépendants.

Compte tenu de ces différences, les stratégies réglementaires devraient privilégier, dans la mesure du possible, des obligations contraignantes au niveau des produits, car elles offrent une meilleure garantie de conformité et un avantage accru pour l'utilisateur. Parallèlement, les autorités de réglementation devraient élaborer et exiger des cadres de transparence et des mécanismes d'audit indépendants afin de responsabiliser les plateformes quant aux modifications algorithmiques apportées à leurs systèmes. Cette double approche reflète les pratiques réglementaires émergentes observées dans le Code de conception adapté à l'âge du Royaume-Uni et le règlement européen sur les services numériques, qui associent des restrictions d'interface visible à des exigences de divulgation des algorithmes et d'accès aux audits.

En résumé : des responsabilités claires et vérifiables en matière de conception de l'interface utilisateur constituent un socle essentiel de sécurité et de protection des utilisateurs, tandis que la gouvernance du système de recommandation nécessite des outils complémentaires de transparence et de supervision pour garantir la conformité et l'efficacité. Lorsque les autorités de réglementation s'intéressent à la conception (paramètres par défaut, objectifs et signaux des systèmes de recommandation, contrôles de la vélocité et de la viralité, et transparence vérifiable), les plateformes apportent des modifications vérifiables qui améliorent la sécurité à grande échelle. L'expérience de l'AADC, de l'application de la DSA et du concept de « sécurité dès la conception » démontre que cette approche est faisable, réplicable et respectueuse des droits.

3.1 Leçons tirées d'autres secteurs d'activité

Depuis longtemps, les sociétés ont recours à la réglementation pour endiguer les risques que les marchés, à eux seuls, ne parviennent pas à maîtriser. Le port de la ceinture de sécurité dans les automobiles, les avertissements sanitaires sur les paquets de cigarettes et les tests de sécurité des médicaments ne sont pas des innovations volontaires, mais des mesures de protection obligatoires, introduites seulement après des années de résistance et de tergiversations de la part de l'industrie. Avant l'introduction de la réglementation sur les ceintures de sécurité et le tabagisme, les sociétés étaient confrontées à des taux alarmants de décès et de blessures évitables, avec d'innombrables vies fauchées par des accidents de la route et des maladies liées au tabagisme qui restaient incontrôlées en raison de l'absence de mesures de sécurité obligatoires et d'interventions de santé publique.[33,34] Ces interventions révèlent une constante : la sécurité publique ne devient une priorité que lorsque les politiques publiques établissent des normes contraignantes pour gérer les risques à l'échelle de la population.

[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.

Les plateformes en ligne se trouvent aujourd'hui à un tournant similaire. Les ajustements volontaires sont largement insuffisants face à l'ampleur des préjudices subis par les groupes vulnérables.[35] Comme pour les voitures, le tabac et les produits pharmaceutiques, les solutions les plus efficaces proviennent de réglementations en amont qui intègrent la sécurité, la transparence et la responsabilité dès la conception.

Une gouvernance proactive et axée sur la conception génère des bénéfices publics tangibles. Les secteurs pharmaceutique et des services financiers, où le risque systémique, la transparence et la surveillance indépendante sont privilégiés sans pour autant imposer les produits finaux, en sont particulièrement concernés. En établissant rapidement des normes contraignantes, les autorités de réglementation peuvent protéger les populations vulnérables, prévenir les préjudices évitables et garantir des progrès sociaux importants sans porter atteinte aux droits fondamentaux.

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

3.2 Positions réglementaires mondiales

Les autorités de réglementation délaissent la surveillance des discours au profit de la correction des défauts de conception. Vous trouverez ci-dessous des exemples, issus de différentes juridictions, illustrant comment elles s'attaquent aux fonctionnalités, aux paramètres par défaut et aux systèmes de recommandation (algorithmes) en amont.

3.2.1 UE — Loi sur les services numériques (DSA)

- **Champ d'application:** Très grandes plateformes en ligne/moteurs de recherche. Approche systémique : évaluation des risques systémiques (art. 34) → atténuation de ces risques par la conception (art. 35) → audits annuels indépendants (art. 37) → choix des recommandations par les utilisateurs (art. 38) → accès réservé aux chercheurs agréés (art. 40). Les mineurs bénéficient d'une protection renforcée (art. 28).

- **Leviers de conception:** Modifications du classement, de la viralité et des paramètres par défaut en fonction des risques ; protection des mineurs (sécurité renforcée, protection de la vie privée dès la conception, art. 28) ; audits annuels (art. 37). Les services doivent exposer aux utilisateurs les principaux paramètres du système de recommandation (art. 27).

- **Systèmes de recommandation:** Proposez au moins une option de non-profilage (art. 38) et indiquez les principaux paramètres sur lesquels les utilisateurs peuvent agir (art. 27). Cela favorise une approche axée sur la qualité et la valeur à long terme, et non pas un simple engagement.

- **Transparence et accès:** Réaliser des évaluations des risques systémiques (art. 34) et adopter des mesures d'atténuation raisonnables et efficaces (art. 35) ; publier des synthèses des risques ; se soumettre à des audits indépendants (art. 37). Garantir l'accès des chercheurs dûment habilités, sous réserve de strictes garanties (art. 40) ; l'UE a adopté un acte délégué précisant les modalités d'accès aux données.

- **Qu'est-ce qui fonctionne?** L'action de la Commission au titre de la loi sur les services numériques (DSA) concernant la fonctionnalité « récompenses pour le visionnage » de TikTok Lite a incité TikTok à la suspendre et, en août 2024, à prendre des engagements contraignants pour supprimer la fonctionnalité dans toute l'UE et ne pas introduire de solution de contournement.36 Il s'agit d'un exemple d'application de la loi sur les risques liés à la conception, ciblant les mécanismes plutôt que le contenu lui-même.

- **Leviers de conception réplicables:** Relier les constats de risques à des correctifs de conception concrets (signaux de classement, contrôles de viralité ou de vitesse, valeurs par défaut) ; centrer les valeurs par défaut protectrices des mineurs ; proposer un flux non profilant que les utilisateurs peuvent choisir ; expliquer clairement les paramètres du système de recommandation ; permettre un accès sécurisé pour les chercheurs et des audits indépendants.

---

[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 Royaume-Uni — Évaluation d'Impact du Code de Conception Adaptée à l'Âge (AADC) & Loi sur la Sécurité en Ligne (OSA)

- **Champ d'application:** L'AADC a entraîné les principaux changements de conception pour protéger les enfants en ligne, sous la supervision de l'ICO. La Loi sur la Sécurité en Ligne (OSA), régulée par Ofcom, étend ces obligations à davantage de services numériques.
- **Leviers de conception:** Confidentialité par défaut, contacts et étiquetages plus sûrs, hygiène des notifications ; obligations systémiques liées au choix de l'utilisateur.
- **Systèmes de recommandation:** Objectifs clairs ; contrôles visibles par l'utilisateur ; tests de mitigation.
- **Transparence et accès:** Journaux de modifications liés aux codes et directives de reporting.[37]
- **Qu'est-ce qui fonctionne?** Les analyses indépendantes montrent 128 changements par défaut effectués par les plateformes (lecture automatique réduite ; alertes nocturnes limitées ; paramètres plus privés par défaut).
- **Leviers de conception réplicables:** Faire en sorte que les paramètres par défaut fassent le travail (intérêt supérieur / droits de l'enfant) ; lier directement les évaluations des risques aux changements de conception.

3.2.3 Australie — eSafety (Sécurité par Conception)

- **Champ d'application:** S'applique à tous les services en ligne accessibles aux utilisateurs en Australie, y compris les plateformes présentant un risque significatif de préjudice ou d'abus, comme les réseaux sociaux, les services de messagerie et les places de marché.
- **Leviers de conception:** Accent particulier mis sur la protection des groupes à risque élevé de violence basée sur le genre facilitée par la technologie (TFGBV).
- **Systèmes de recommandation:** Divulgation objective, choix utilisateur (chronologique et selon les intérêts uniquement), tests continus.
- **Transparence et accès:** Maintenir un journal de modifications clair et à jour documentant les modifications système liées à la sécurité des utilisateurs.
- **Qu'est-ce qui fonctionne?** Correctifs documentés après engagement ; explications plus claires des recommandations dans les centres d'aide.[38]
- **Leviers de conception réplicables:** Publier un pack de modèles (SRA, audit, journal de modifications) ; échelle transparente pour corriger la conception avant sanctions.

---

[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 Etats-Unis — New York (SAFE for Kids Act et Child Data Protection Act) & Vermont (Act 63, 2025 AADC)

- **Champ d'application:** Services utilisés par des mineurs (<18 ans) à New York et Vermont, incluant flux, notifications et traitement des données enfants.
- **Leviers de conception:** Restriction des mécanismes addictifs ou à forte interaction sans consentement, limites de notifications nocturnes, minimisation de la collecte et usage des données enfants. L'Act 63 du Vermont ajoute des standards d'assurance de l'âge via des règles du procureur général.
- **Systèmes de recommandation:** Curation non addictive pour les mineurs, objectifs transparents, contrôles simples ; optimisation uniquement basée sur l'engagement découragée.
- **Transparence et accès:** Documentation et supervision claires obligatoires (DPIA/DIA) ; le procureur général de New York peut émettre des directives et demander des rapports périodiques.
- **Qu'est-ce qui fonctionne?** Les plateformes mettent en place des contrôles de flux et limitent les notifications pour les mineurs ; au Vermont, premières avancées vers un consentement transparent et des paramètres adaptés à l'âge.
- **Leviers de conception réplicables:** Considérer flux et notifications comme obligations de conception ; intégrer assurance d'âge et minimisation des données enfants dans le produit ; relier paramètres par défaut aux droits et sécurité des enfants.

3.2.5 États-Unis — Minnesota (HF 4400 et Statut 325M.33)

- **Champ d'application:** Plateformes de médias sociaux exerçant des activités commerciales au Minnesota ou desservant des résidents du Minnesota ; dispositions relatives à la transparence en vigueur à compter du 1er juillet 2025.

- **Leviers de conception:** Mettre l'accent sur la transparence de la conception concernant les mécanismes qui causent des dommages : limites d'interaction/de vitesse publiées, signaux de qualité et de préférence, pratiques de notification (y compris les statistiques nocturnes) et résumés des expériences sur les produits.

- **Systèmes de recommandation:** Explication publique de la manière dont les systèmes de classement utilisent les signaux de qualité et de préférence exprimée et leurs pondérations relatives (publiée sur le site web).

- **Transparence et accès:** Exigences légales de publication de la « transparence algorithmique » (§ 325M.33) ; Le résumé de la recherche de la Chambre indique l'application par le procureur général et la date d'entrée en vigueur du 1er juillet 2025.

- **Qu'est-ce qui fonctionne?** De nouveaux tableaux de bord/informations publiques sont attendus (par exemple, des statistiques en percentile sur les interactions des utilisateurs ; le nombre de notifications nocturnes ; des descriptions d'expériences).

- **Leviers de conception réplicables:** Imposer des publications publiques et lisibles par machine sur les signaux, les pondérations, les limites, les notifications et les expériences ; lier l'application de ces règles à des objectifs de recommandation transparents et compréhensibles par l'utilisateur.

3.2.6 Brésil — Statut numérique pour les enfants et adolescents (ECA Digital)

- **Champ d'application:** L'ECA Digital s'applique à tout service internet, application ou produit technologique destiné ou susceptible d'être utilisé par des mineurs (<18 ans).

- **Leviers de conception:** Exigences strictes pour la protection des jeunes : vérification de l'âge, contrôles parentaux clairs, restrictions sur la publicité ciblée.

- **Systèmes de recommandation:** Documentation des objectifs des algorithmes, limitation du profilage des jeunes ; choix de conception favorisant la valeur à long terme ; données des mineurs protégées contre les intrusions de vie privée.

- **Transparence et accès:** Rapports publics semestriels sur les mesures de protection des jeunes ; justification claire des suppressions de contenu.

- **Qu'est-ce qui fonctionne?** Les initiatives pilotes de protection des jeunes ont conduit les plateformes à mettre en place la vérification d'âge avant l'entrée en vigueur de la loi en mars 2026.

- **Leviers de conception réplicables:** Ajouter divulgation des recommandations et paramètres par défaut jeunesse dans la base de droits ; permettre un accès progressif aux chercheurs.

3.2.7 Indonésie — Réglementation PP TUNAS n° 17 de 2025

- **Champ d'application:** PP TUNAS fixe les règles techniques et opérationnelles pour les opérateurs de systèmes électroniques en Indonésie afin de protéger les données personnelles des enfants, en complément de la loi générale sur la protection des données (PDP).
- **Leviers de conception:** Conception centrée sur l'enfant ; protections par défaut contre les designs manipulateurs pour les mineurs (phase initiale).
- **Systèmes de recommandation:** Encouragement de modes plus sûrs : flux chronologiques ou sans profilage, réduction de l'exposition à contenu nocif et aux éléments addictifs (lecture automatique, défilement infini).
- **Transparence et accès:** Orientation vers l'évaluation des risques liés à la conception ; mécanismes d'accès en développement.
- **Qu'est-ce qui fonctionne?** Déclarations politiques signalant un passage de la modération pure à des mesures de protection centrées sur la conception.
- **Leviers de conception réplicables:** Codifier les paramètres par défaut pour les mineurs (confidentialité élevée) ; piloter un flux chronologique et sans profilage avec journaux de modifications publics.

3.2.8 Corée — Loi de révision sur la protection (Loi "Cendrillon")

- **Champ d'application:** Jeux en ligne pour mineurs (<16 ans) en Corée du Sud, initialement limités la nuit pour protéger la jeunesse.[39]
- **Leviers de conception:** Interdiction de jouer entre 0h et 6h ; remplacée en 2021 par un système flexible de "heures de jeu sélectives" permettant le contrôle parental.[40]
- **Systèmes de recommandation:** Encouragement à éviter les boucles addictives et à promouvoir des expériences sûres pour les jeunes.
- **Transparence et accès:** Vérification obligatoire de l'âge et reporting par les fournisseurs de services.
- **Qu'est-ce qui fonctionne?** Réduction du jeu nocturne, mais critiques sur le contournement par les mineurs ; la réforme de 2021 donne plus de contrôle aux familles.[41]
- **Leviers de conception réplicables:** Contrôle parental et exemptions ; systèmes de vérification de l'âge ; modération algorithmique ciblant le contenu addictif.

---

[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 S'appuyer sur la dynamique africaine

Pourquoi le contexte est-il si propice en Afrique? Au cours de la dernière décennie, les pays africains ont mis en place des infrastructures solides : programmes de lutte contre la cybercriminalité, protection des données et régulation des médias, sans oublier les investissements publics dans l'éducation numérique et la vérification des faits. Ces progrès permettent désormais de privilégier des règles de conception favorisant les comportements sociaux (flux d'actualité, paramètres par défaut, contrôle de la viralité) plutôt que de se reposer uniquement sur le retrait de contenus ou la sensibilisation des utilisateurs.

La Stratégie continentale de l'Union africaine en matière d'intelligence artificielle fournit un cadre fondamental pour atténuer les préjudices en ligne et promouvoir une conception sûre de l'IA sur le continent.[42] Elle préconise des cadres réglementaires complets qui privilégient la réduction des risques et alignent explicitement la politique en matière d'IA sur les principes relatifs aux droits humains. Cette stratégie établit des protections juridiques et plaide pour une conception sûre et transparente des systèmes d'IA.

Voici points saillants de la réglementation au niveau national au sein de l'Union africaine :

- **Kenya:** La loi sur l'utilisation abusive des ordinateurs et la cybercriminalité (2018) criminalise les fausses publications et la désinformation avec des peines allant jusqu'à 10 ans d'emprisonnement ou des amendes, s'attaquant à la fois à la désinformation intentionnelle et à ses dommages publics.

- **Côte d'Ivoire:** La Côte d'Ivoire a renforcé la protection des données avec la loi n° 2013-450, appliquée par l'ARTCI, qui garantit des droits tels que l'accès aux données, leur rectification et l'opposition à leur traitement. Le pays lutte également contre les fausses informations et la désinformation grâce à des programmes d'éducation aux médias comme Désinfox Côte d'Ivoire, qui forment journalistes et jeunes à la vérification de l'information et à la réduction de l'impact de la désinformation sur la cohésion sociale.

- **Niger:** Le Haut Conseil de la Communication (CCC) du Niger supervise la réglementation de l'information audiovisuelle et en ligne afin de garantir le respect des normes médiatiques et le pluralisme. Le pays participe activement aux efforts menés par la CEDEAO pour harmoniser la législation et les politiques de lutte contre la cybercriminalité en Afrique de l'Ouest. Par ailleurs, le Niger a mis en place une commission multipartite axée sur l'intégrité de l'information, qui encourage la collaboration entre les acteurs gouvernementaux, la société civile et le secteur privé afin de lutter contre la désinformation et de renforcer la confiance numérique.

- **Sénégal:** Autorité de protection des données (CDP) forte et longue tradition de contrôle des médias ; débats sur la désinformation associés à un soutien gouvernemental à l'éducation aux médias et à l'information et au renforcement des capacités des rédactions.

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

- **Mali:** L'Autorité de Protection des Données à Caractère Personnel (APDP) du Mali, créée par la loi n° 2013-015 et opérationnelle depuis 2016, fait office d'autorité nationale de protection des données. Elle veille à un équilibre entre sécurité et droits par le biais de mesures d'application et d'une collaboration régionale, notamment sa participation au Forum de Bamako sur la cohésion numérique et sociale.

- **Nigéria:** La loi nigériane sur la protection des données (NDPA 2023) s'appuie sur la loi fondamentale NDPR (2019) en établissant un cadre juridique complet pour la protection des données, avec une application renforcée, un contrôle accru pour l'utilisateur et des obligations spécifiques au secteur.[43,44] La NDPA impose l'enregistrement, les audits de conformité et la désignation de responsables de la protection des données, et son application est déjà en cours par la Commission nigériane de protection des données (NDPC).

Engagements volontaires des plateformes et des régulateurs. La Déclaration d'Abidjan engage les plateformes et les régulateurs à : assurer une protection renforcée des mineurs ; promouvoir des systèmes de recommandation privilégiant la diversité des sources ; garantir une transparence concrète et l'accès des chercheurs ; et mettre en place un canal de coopération régionale pour assurer le suivi de la mise en œuvre.45

Considérez les initiatives multipartites suivantes:

- Le Cadre politique pour l'intégrité de l'information en Afrique de l'Ouest et au Sahel, adopté à la suite de la Conférence régionale sur l'intégrité de l'information en Afrique de l'Ouest et au Sahel, enjoint aux autorités de régulation de « se concentrer sur les conditions systémiques qui permettent aux contenus nuisibles et trompeurs de prospérer » et adopte une approche systémique en amont. Il constate que les mesures antérieures « n'ont souvent pas permis de s'attaquer aux causes systémiques de la désinformation et des discours de haine ». Il préconise une transparence et une responsabilité accrues des plateformes, notamment par la réalisation d'évaluations régulières et indépendantes des risques algorithmiques.[46]

- Le Forum de Bamako sur la Cohésion Numérique et Sociale (depuis 2023) réunit le Burkina Faso, le Mali et le Niger, ainsi que des représentants des médias, des plateformes et de la société civile ; sa Déclaration de 2024 se concentre sur les aspects fondamentaux – paramètres par défaut adaptés aux enfants, limites de la lecture automatique et du défilement infini, objectifs plus clairs des systèmes de recommandation – tout en soutenant l'éducation numérique et la vérification des faits.[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.

- La Coalition nationale pour la liberté d'expression et la modération des contenus (FeCoMo) au Kenya mobilise les partenariats entre le gouvernement, la société civile et le monde universitaire pour promouvoir un espace numérique sain et sûr.

- En septembre 2025, des membres du Forum de Bamako et de FeCoMo ont publié la « Déclaration de Nairobi » appelant à une responsabilisation des plateformes dès la conception du système.[49]

- La communauté de pratique Techsocietal African Online Safety a régulièrement fait progresser les approches de sécurité intégrée dès la conception lors des événements organisés pour ses membres.[50]

- Build Up mène des recherches émergentes sur les préjudices, la division et la polarisation afin de mesurer « l'empreinte de polarisation » des plateformes de médias sociaux, dans le but de comprendre leur coût pour la société.[51]

S'appuyant sur cette dynamique, les autorités et la société civile peuvent inviter les plateformes à tester des modèles prosociaux volontaires (invites, contrôles de diversité d'exposition, invites de pause de session) à travers l'Afrique, la société civile et les chercheurs évaluant conjointement les résultats pour les utilisateurs de plateformes numériques selon les indicateurs proposés dans la 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. Définitions et clauses juridiques

Cette bibliothèque propose un lexique de conception partagé et des extraits de clauses prêts à l'emploi tirés des lois existantes. Ces éléments offrent des règles pratiques, claires, applicables et neutres quant au contenu, susceptibles d'inspirer les réglementations futures.

4.1 Définitions — Lexique du design

Interface médiatisée par l'IA — Toute fonctionnalité qui génère ou réécrit de manière significative le contenu ou les interactions pour l'utilisateur (par exemple, chatbot, réponse automatique, récapitulatif).

Mécanismes addictifs — Des schémas qui prolongent l'utilisation sans intention explicite de l'utilisateur (par exemple, le défilement infini, la lecture automatique suivante), en particulier pour les mineurs.

Backend — Processus algorithmiques et système qui se déroulent « en coulisses », notamment les calculs du système de recommandation, les ajustements de classement, la pondération des signaux et les décisions d'amplification du contenu.

Évaluation de l'impact spécifique à l'enfant (EISE) — Une évaluation appliquant la norme « dans le meilleur intérêt de l'enfant » aux effets prévisibles sur les mineurs.

Journal des modifications — Un registre de toutes les mises à jour, corrections et améliorations importantes apportées à un projet, un produit ou un logiciel au fil du temps.

Dark Pattern — Une interface qui subvertit matériellement l'autonomie de l'utilisateur (tromperie, obstruction, choix forcé).

Par défaut — La configuration prédéfinie affichée lorsque l'utilisateur ne fait rien.

Interface utilisateur — Éléments d'interface visibles et destinés à l'utilisateur, avec lesquels il peut interagir directement, tels que les paramètres de lecture automatique, les commandes de notification, les rappels de pause et les paramètres de confidentialité par défaut.

Contenu généré par IA (médias synthétiques) — Texte, image, audio ou vidéo produit par un modèle automatisé, non enregistré à partir du monde réel.

Mode Valeur à long terme (LTV) — Un paramètre de flux sélectionnable par l'utilisateur qui privilégie la crédibilité, la qualité, la diversité des sources et le bien-être par rapport à l'engagement à court terme.

Modification de la conception matérielle — Une modification des objectifs/signaux/pondérations du système de recommandation ; des paramètres par défaut de base (confidentialité/notifications/contact) ; des contrôles de viralité ou de vitesse ; ou des expériences connexes.

Mode sans profilage — Une option de flux qui ne repose pas sur le profilage comportemental.

Provenance du contenu — Méthode permettant de documenter et de vérifier l'origine, l'historique et les modifications d'un contenu numérique, tel qu'une image, une vidéo, un fichier audio ou un document.

Limitation du débit — Un plafond documenté pour les actions à haute vélocité (publication, partage, étiquetage, invitations).

Système de recommandation — Tout système automatisé de classement, de personnalisation ou de suggestion de contenu ou de contacts.

Évaluation des risques systémiques ou de l'impact sur la conception (SRA/DIA) — Une analyse préalable au lancement des risques prévisibles liés à une modification de conception et des mesures d'atténuation choisies.

Violence fondée sur le genre facilitée par la technologie (TFGBV) — Tout acte de préjudice (physique, psychologique, social ou économique) commis, aidé, amplifié ou aggravé par le biais des technologies numériques ou des plateformes en ligne et dirigé contre des personnes en raison de leur sexe.

4.2 Leviers de changement que les régulateurs utilisent déjà

CHAQUE ÉLÉMENT = PROBLÈME QU'IL RÉSOUT → COURT EXTRAIT TEXTUEL → ÉTIQUETTE SOURCE

4.2.1 Normes de conception et valeurs par défaut

Limitez par défaut qui peut trouver votre compte.
« Par défaut, le compte d'un utilisateur est configuré de manière à ce qu'il ne soit accessible qu'aux contacts invités de l'utilisateur ou avec son consentement. » Source : Minnesota HF 4400, Subd. 3 (« Paramètres de confidentialité par défaut »)

Limiter la visibilité de la géolocalisation par défaut.
« Par défaut… limiter au strict minimum la capacité des autres titulaires de compte à connaître ou à partager la géolocalisation d'un utilisateur. » Source : Minnesota HF 4400, alinéa 3(b)

Couvre-feu nocturne pour les mineurs.
« Il est interdit d'envoyer sciemment une notification push… à un mineur pendant la nuit, sauf si l'utilisateur a donné son accord en précisant une heure de début et de fin. » Source : Loi SAFE for Kids de l'État de New York (S7694-A)

Paramètres de confidentialité élevés pour les enfants par défaut.
« Les paramètres doivent être configurés par défaut sur un niveau de confidentialité élevé. » Source : Code de conception adapté à l'âge du Royaume-Uni (ICO), norme 7

Laissez les enfants dire « ne me montrez pas ça » (rétroaction négative).
« Permettre aux enfants de donner leur avis négatif sur les contenus recommandés » Source : Loi britannique sur la sécurité en ligne – Code de bonnes pratiques pour les enfants (RS3)

Pas de « contenus addictifs » pour les mineurs sans contrôle.
« Il est illégal de fournir un contenu addictif à un utilisateur concerné, sauf si (A) l'utilisateur n'est pas mineur ; ou (B) il existe un consentement parental vérifiable. » Source : Loi SAFE for Kids de l'État de New York (S7694-A)

Limiter les comportements à haute vélocité (lutte contre le spam et le brigading).
« Doit fixer et faire respecter des limites quotidiennes strictes et raisonnables concernant… les publications ou les republications… les nouveaux comptes suivis… la fréquence des identifications… et autres actions susceptibles de causer un préjudice par abus. » Source : Minnesota HF 4400, Subd. 2

CHAQUE ÉLÉMENT = PROBLÈME QU'IL RÉSOUT → COURT EXTRAIT TEXTUEL → ÉTIQUETTE SOURCE

4.2.2 Gouvernance des systèmes de recommandation — Meilleurs flux

Proposez une option de flux RSS sans profilage.
« Les destinataires du service doivent disposer d'au moins une option de système de recommandation non fondé sur le profilage. » Source : Règlement (UE) sur les services numériques, article 38

Optimisez la qualité et les préférences exprimées (et non l'engagement brut).
« Le système de classement algorithmique doit optimiser le contenu… qui (1) est jugé de haute qualité par un ensemble varié de titulaires de comptes ; et (2) correspond aux préférences exprimées par l'utilisateur. » Source : Minnesota HF 4400, Subd. 1

Réduire le ciblage des jeunes par le biais de suggestions.
« S'abstenir d'utiliser des groupes, comptes, utilisateurs, services, publications et produits ciblés ou suggérés sur le compte jeune. » Source : Loi S385 du Wisconsin sur les comptes jeunes

Réduisez la visibilité et évitez de recommander des contenus nuisibles aux enfants.
« Veiller à ce que les contenus susceptibles d'être [primitivement préjudiciables] ne soient pas recommandés aux enfants » et « soit moins visibles dans les flux de recommandations destinés aux enfants. » Source : Loi britannique sur la sécurité en ligne – Code de bonnes pratiques pour les enfants (RS1, RS2)

4.2.3 Transparence et accès des chercheurs

Publications publiques en langage clair sur les algorithmes (signaux et pondérations).
« Maintenir une page publique expliquant… les signaux de classement et leur importance relative »
Source : Minnesota Stat. 325M.33 (HF 4400)

Évaluations des risques et mesures d'atténuation des risques liées à la conception.
« Réaliser des évaluations des risques… et mettre en place des mesures d'atténuation raisonnables, proportionnées et efficaces » Source : Règlement sur les services numériques de l'UE, articles 34 et 35

Réaliser des audits indépendants.
« Les très grandes plateformes seront soumises à des audits indépendants. » Source : Règlement européen sur les services numériques, articles 37 à 39

Garantir l'accès aux chercheurs agréés.
« Offrir aux chercheurs agréés un accès aux données… pour des recherches contribuant à la détection, à l'identification et à la compréhension des risques systémiques »
Source : Règlement européen sur les services numériques, article 40

4.3 Réduction des risques par la réglementation des conceptions

Expérience utilisateur (préjudice)	Mécanicien concepteur (conducteur)	Remède dans la loi ou le code	À quoi ressemble le succès ?
« Mon fil est rempli de contenus de faible qualité ou trompeurs. »	Classement basé uniquement sur l'engagement ; partages en un clic ; tendances guidées par la vitesse	EU DSA Art. 38 option non-profilage ; MN HF 4400 Subd.1 optimisation pour qualité & préférences exprimées	↑ source crédible ; ↑ diversité de l'exposition ; ↓ vitesse de partage ; ↓ clics regrettables
« Je continue d'être rappelé tard la nuit. »	Notifications push nocturnes	NY SAFE (S7694-A) limites de push nocturnes pour mineurs	↓ notifications nocturnes ; ↑ qualité du sommeil signalée chez les mineurs
« Je n'arrive pas à arrêter de défiler »	Défilement infini & lecture automatique	NY SAFE (S7694-A) restriction des flux addictifs pour mineurs ; UK AADC paramètres haute confidentialité / sécurité enfants	↓ durée de la session ↑ utilisation de la pagination ou des invites de pause
« Des inconnus continuent de me taguer ou m'envoyer des messages. »	Messages privés ouverts ; tags massifs ; limitations faibles pour nouveaux comptes	MN HF 4400 Subd.2 limites quotidiennes & limites plus strictes pour nouveaux comptes	↓ messages privés non sollicités ; ↓ harcèlement de nouveaux comptes durant la première semaine
« Je vois du contenu nuisible recommandé à mon enfant. »	Exposition des jeunes via les recommandations	UK OSA code enfants (RS1/RS2) ne pas recommander / réduire la visibilité ; DSA Art. 38 non-profilage	↓ exposition des enfants aux catégories signalées ↑ utilisation efficace des commentaires négatifs
« Je ne trouve pas ou ne conserve pas les paramètres que je veux. »	UX trompeuse ; paramètres basse confidentialité par défaut	UK AADC "haute confidentialité par défaut" ; MN HF 4400 Subd.3 découvrabilité / limites géolocalisation par défaut	↑ réglage de la fréquence d'utilisation ; ↑ contrôle perçu
« Le spam / brigading envahit les commentaires. »	Publication rapide, partages et tags massifs	MN HF 4400 Subd.2 limites quotidiennes	↓ pics de signalements venant de nouveaux comptes ; ↓ harcèlement coordonné
« Personne ne peut vérifier les affirmations de la plateforme. »	Systèmes opaques ; pas d'accès	DSA Arts. 34–35, 39–40 évaluations des risques, audits, accès chercheurs vérifiés	Journaux de modifications publics ; résultats d'audits indépendants ; recherche publiée et réplicable

4.4 L’expérience utilisateur comme indicateur de succès

Le véritable critère d’efficacité des correctifs de conception est l’amélioration réelle de l’expérience en ligne quotidienne des utilisateurs, et non pas seulement ce qu’indiquent les tableaux de bord. Les principales plateformes sociales et les organismes de réglementation adoptent déjà cette approche en menant régulièrement des enquêtes sur l’expérience utilisateur.[52]

- Snap Inc. réalise chaque année une étude sur l'indice de bien-être numérique auprès d'adolescents et de jeunes adultes dans six pays (portant sur l'ensemble de leurs activités en ligne, et pas seulement sur Snapchat) afin de mesurer le bien-être psychologique de la génération Z.[53]
- L’Ofcom, l’autorité britannique de régulation des communications, mène chaque année une enquête intitulée « Expérience des internautes face aux préjudices en ligne » afin de quantifier la fréquence à laquelle le public est confronté à des contenus préjudiciables, à des atteintes à la vie privée ou à des menaces à la sécurité.[54]
- En Australie, l’enquête 2024 de la Commission de la sécurité en ligne intitulée « Protéger les enfants en ligne », menée auprès de 3 454 enfants, a révélé que 74 % des enfants âgés de 10 à 17 ans avaient été exposés à des contenus préjudiciables et que plus de la moitié avaient subi du cyberharcèlement, soulignant ainsi la nécessité de renforcer les mesures de protection dès la conception.[55]

Parallèlement, des chercheurs indépendants ont mis au point des indices standardisés. Par exemple, l’indice USC Neely des médias sociaux utilise des enquêtes continues pour suivre l’évolution des expériences positives et négatives des utilisateurs sur chaque plateforme.[56,57] Les plateformes elles-mêmes recueillent également ces données (par exemple, l’enquête interne d’Instagram sur les mauvaises expériences et les rencontres négatives), mais ces résultats sont rarement publiés.[58]

La solution consiste à rendre ce processus de rétroaction public, par exemple en créant un « Observatoire de l’expérience numérique » régional et multipartite, chargé de gérer un panel d’utilisateurs indépendant et continu, ainsi que des enquêtes d’opinion régulières intégrées aux applications. Les plateformes participent, les autorités de régulation se réunissent, la société civile cogère et les chercheurs conçoivent les méthodes et gèrent les ensembles de données publics.

---

[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.

Que mesurer? Voici 10 exemples de questions simples qu'une enquête centrée sur l'utilisateur pourrait poser régulièrement – abordant tous les aspects, de la qualité du contenu à la sécurité personnelle, afin de déterminer si les récentes modifications de conception ont un réel impact :

- **Temps bien employé:** « Après avoir utilisé [l'appli] aujourd'hui, vous sentez-vous mieux ou moins bien qu'avant ? »
- **Crédibilité du contenu:** « À quelle fréquence voyez-vous des publications que vous considérez comme bien sourcées ? »
- **Relier les points de vue sur les contenus:** « Au cours de la semaine écoulée, avez-vous constaté des points de vue variés sur les sujets que vous suivez ? »
- **Les recommandations correspondent-elles à vos attentes à long terme?** « Votre flux reflète-t-il ce que vous avez indiqué vouloir ? »
- **Des paramètres qui restent en place:** « Lorsque vous modifiez les paramètres (notifications, personnes autorisées à vous envoyer des messages privés), restent-ils tels que vous les avez définis ? »
- **Perturbations du sommeil:** « Avez-vous été réveillé(e) par des notifications tard dans la nuit cette semaine ? »
- **Contacts indésirables:** « Combien de fois cette semaine une personne que vous ne connaissez pas vous a-t-elle envoyé un message privé ou vous a-t-elle identifié(e) ? »
- **Harcèlement ou usurpation d'identité:** « Cette semaine, avez-vous subi ou été témoin de harcèlement ou d'usurpation d'identité sur [app] ? » (Vécu / Témoin / Les deux / Aucun) → Si vous avez subi : « Avez-vous facilement pu le signaler et obtenir de l'aide ? »
- **Précision et transparence de l'IA (sous condition):** « Si vous avez utilisé un chatbot/une recherche basée sur l'IA, ses réponses provenaient-elles de sources auxquelles vous faites confiance ? »
- **Clarté des médias synthétiques (conditionnelle):** « Pourriez-vous facilement déterminer si une image, une vidéo ou un texte a été généré/modifié par une IA ? »

Les organismes de réglementation ne devraient pas avoir à deviner l'efficacité d'une nouvelle règle de sécurité ou d'une modification de conception : les utilisateurs pourront le leur dire, si on les interroge. Ce type de panel d'expérience utilisateur continu et indépendant, associé à de simples sondages intégrés à l'application, est pratique et accessible. Cette boucle de rétroaction garantit la transparence : si les chiffres n'évoluent pas dans la bonne direction, les organismes de réglementation et les plateformes sauront qu'il est temps de rectifier le tir.

Les enquêtes auprès des utilisateurs peuvent être complétées par d'autres indicateurs et mesures s'appuyant sur des normes et évaluations reconnues. Cela inclut la conformité aux normes reconnues (telles que la norme IEEE 2089 : Norme pour un cadre de services numériques adaptés à l'âge et fondé sur les 5 principes des droits de l'enfant), évaluée par des audits d'accréditation indépendants réalisés par des tiers.[59] De plus, les évaluations menées par des experts externes – portant sur les interfaces en ligne, les systèmes, les paramètres, les outils, les fonctionnalités, ainsi que les mécanismes de signalement, de retour d'information et de traitement des réclamations – offrent un niveau de responsabilité supplémentaire.

Pour que ces indicateurs d'expérience utilisateur et autres mesures contribuent réellement à la responsabilisation, un observatoire pourrait inclure des régulateurs, des universitaires, des représentants de la société civile et des conseillers jeunesse afin de refléter une diversité de points de vue. Les plateformes bénéficieraient d'une protection juridique (« sphère de sécurité ») lors du partage de données avec l'observatoire, et des chercheurs accrédités pourraient obtenir un accès aux données respectueux de la vie privée pour approfondir l'analyse des résultats.

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

5. Conclusion

Ce guide pratique part d'un constat simple : les préjudices en ligne ne sont pas inévitables, mais la conséquence prévisible des choix de produits. Lorsque les autorités de réglementation se concentrent sur les mécanismes de conception plutôt que sur la surveillance des opinions, la sécurité devient mesurable, respectueuse des droits et applicable à grande échelle. L'expérience acquise grâce aux dispositifs de protection de l'enfance, à l'application de la loi sur la sécurité des plateformes (DSA) et à l'approche « sécurité dès la conception » démontre que des règles claires et vérifiables concernant les flux, les paramètres par défaut, les notifications, le balisage et la transparence entraînent de véritables modifications des produits, notamment pour les enfants et les adolescents, sans pour autant restreindre la liberté d'expression. Ces modifications réduisent la diffusion de fausses informations et de contenus clivants, et diminuent les incitations à publier des contenus nuisibles et sources de division. La voie à suivre est concrète et immédiate : les plateformes doivent :

1. Mettez fin aux pratiques de manipulation et freinez les mécanismes addictifs.

2. Rendre les services sécurisés par défaut pour les mineurs.

3. Proposer de meilleures options de flux, optimisées pour une valeur à long terme en termes de crédibilité, de contenu de liaison et de préférences des utilisateurs.

4. Tester les modifications matérielles avant le lancement, publier des journaux de modifications en langage clair et soumettre à des audits indépendants et à un accès de recherche respectueux de la vie privée.

5. Mesurer le succès par les expériences des utilisateurs, renforçant ainsi l'engagement de multiples parties prenantes dans la compréhension des défis et la mesure conjointe des progrès.

Appliquez les mêmes obligations aux assistants IA et aux fonctionnalités génératives. Privilégiez des exigences claires et visibles et associez-les à une transparence et à des obligations d'audit rigoureuses pour la gouvernance des systèmes de recommandation, ainsi qu'à des indicateurs d'expérience utilisateur. De nombreux pays sont mûrs pour une approche de gouvernance intégrée dès la conception, s'appuyant non seulement sur la dynamique africaine, mais aussi sur un consensus mondial croissant.

Cette approche s’appuie sur des cadres et politiques internationaux tels que les Principes directeurs de l’UNESCO pour la gouvernance des plateformes numériques, l’Observation générale n° 25 de la Convention relative aux droits de l’enfant des Nations Unies, l’AADC du Royaume-Uni, le programme australien de sécurité intégrée dès la conception et le cadre politique de Praia adopté lors de la Conférence régionale de septembre 2025 au Cap-Vert. Ces politiques et réglementations préconisent une coopération multipartite et une attention particulière aux préjudices systémiques liés à la conception. Surtout, ces cadres anticipent et permettent **une collaboration constructive avec la société civile et les chercheurs,** qui sont prêts à coévaluer les résultats afin que les plateformes puissent démontrer, et non simplement affirmer, que les modifications apportées à la conception améliorent l’expérience utilisateur au quotidien.

Si les autorités de régulation définissent ces obligations dès maintenant, les plateformes s'orienteront naturellement vers le bien-être, l'autonomie et la cohésion sociale. C'est en préparant le terrain pour de meilleurs résultats – plutôt qu'en modérant a posteriori – que nous protégerons les droits, réduirons les préjudices à grande échelle et rétablirons la confiance dans l'espace public numérique.

6. Annexe : Preuves des préjudices numériques en Afrique

À travers l'Afrique, huit années d'études évaluées par des pairs, de rapports d'ONG et d'évaluations des forces de l'ordre révèlent que les méfaits des plateformes numériques sont répandus et récurrents. Parmi les principaux problèmes au quotidien : l'utilisation excessive ou compulsive, les risques pour les enfants, les violences sexistes facilitées par les technologies et les escroqueries, souvent amplifiés par des choix de conception tels que les flux optimisés pour l'engagement, le défilement infini ou la lecture automatique, les messages privés ouverts et l'absence d'outils intégrés de vérification des faits ou du contenu.

- **Dépendance :** Une revue systématique et une méta-analyse de 22 études africaines réalisées en 2022 ont révélé une prévalence globale de la dépendance à Internet d'environ 40 % (plus élevée dans les cohortes universitaires), soulignant que l'usage intensif et difficile à contrôler est courant sur tout le continent – et non un problème « exclusivement occidental ».[60]

- **Violences faites aux femmes et aux filles (VFFF) :** L'étude féministe africaine « Alternate Realities, Alternate Internets » documente les violences en ligne courantes ; dans son échantillon, 28 % des femmes ont déclaré avoir subi des violences sexistes et sexuelles et 71 % des incidents signalés se sont produits sur Facebook (WhatsApp/Twitter/X et Instagram étant également fréquemment utilisés).[61] L'étude mondiale de Plan International (incluant les pays africains) a révélé que 58 % des filles étaient victimes de harcèlement en ligne, ce qui les exclut souvent du débat public.[62]

- **Escroqueries et fraudes :** INTERPOL identifie les escroqueries en ligne comme la principale cybermenace signalée par les pays membres africains, notamment l'hameçonnage/l'ingénierie sociale, les faux systèmes d'investissement et la fraude via les plateformes qui exploitent à grande échelle les réseaux sociaux et les applications de messagerie.[63]

- **Discours de haine et polarisation :** le classement basé sur l'engagement et les partages rapides intensifient les discours incendiaires et limitent l'accès des utilisateurs à la diversité des points de vue. Ces pics coïncident souvent avec des élections, des crises ou des violences communautaires. Au Sahel, des observateurs ont signalé des recrudescences de contenus anti-minorités (notamment des propos anti-Fulanis) lors des pics de conflit. Au Kenya, la famille du professeur éthiopien Meareg Amare Abrha a intenté une action en justice historique, alléguant que les manquements de Facebook en matière de lutte contre la haine et l'incitation à la violence ont contribué à son assassinat en 2021. Cette action a été autorisée à se poursuivre en 2025, soulignant ainsi comment les défaillances de Facebook en matière de diffusion de contenus haineux peuvent avoir des conséquences concrètes.[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.

À propos du Conseil sur la technologie et la cohésion sociale

Créé en 2023, le Conseil sur la Technologie et la Cohésion Sociale réunit des technologues, des artisans de la paix, des universitaires et des influenceurs de la politique autour d'une question fondamentale : et si la technologie pouvait favoriser la confiance et la collaboration au lieu d'alimenter la polarisation et la violence ?

Avec plus de 70 membres répartis dans 22 pays, le réseau intersectoriel et mondial du Conseil défend des politiques technologiques prosociales, mesure l'impact de la technologie sur la cohésion sociale et met en avant les preuves de produits et de pratiques qui favorisent des sociétés sûres, saines et cohésives.

En mettant l'accent sur le design comme levier essentiel du changement, le Conseil fait le lien entre la recherche, les politiques et l'innovation, démontrant comment la technologie peut renforcer la confiance, préserver la dignité humaine et devenir une force de paix et de collaboration.