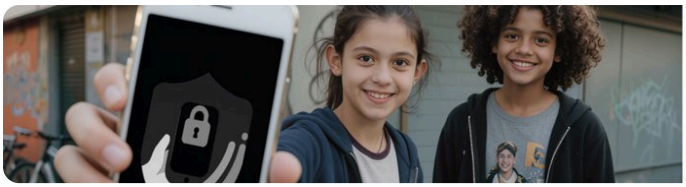


Note méthodologique, scientifique et éthique

Association SOS Écrans — Shieldy · PREVENTO · PREVENTO GO

> *Nous ne cherchons pas à convaincre, mais à exposer une méthode. Nous décrivons ce que nous faisons, sur quoi nous nous appuyons, et — avec autant de soin — ce que nous ne savons pas encore. Chaque affirmation porte sa source ou son statut. Quand une donnée manque, nous le disons plutôt que de la combler.*

Version 1 — mai 2026 · Document destiné aux experts et aux encadrants
Contact : contact@shieldy.org

1. Résumé exécutif

Qui, quoi, pourquoi

SOS Écrans est une association loi 1901 dont l'objet, déclaré aux statuts, est de *prévenir et sensibiliser aux risques liés aux écrans et au cyberharcèlement chez les mineurs ; accompagner les familles, les enseignants et les éducateurs par des outils numériques gratuits*.

Elle porte un écosystème de trois briques :

- **PREVENTO** — un site web de prévention en classe : un atelier-quiz situationnel piloté par l'enseignant, conçu pour s'enseigner sur plusieurs tranches d'âge (6–18 ans). Gratuit, sans compte, sans installation, anonyme.
- **PREVENTO GO** — une *simulation sociale* : ni jeu vidéo, ni quiz, ni film. Un micro-récit jouable, à la première personne d'un **témoin ordinaire**, qui *montre la mécanique* d'une exclusion de groupe (les silences, le « vu à 23 h », les micro-validations) sans insulte, sans score, sans morale. Le vrai produit, c'est le débrief qui suit. C'est un concept abouti, fondé sur un inventaire mondial de la prévention et une doctrine de conception ; le prototype reste à construire.
- **Shieldy** — une application mobile (priorité **Android**) qui repère, dans le texte qui transite par le téléphone de l'enfant (écrit *et* reçu), des signaux de risque (cyberharcèlement, grooming, sextorsion) et alerte le parent sur le *niveau* et la *nature* du risque, jamais sur le contenu ni l'identité d'un tiers. Sur Android, **l'analyse du texte se fait sur l'appareil** — il ne part pas vers nos serveurs.

La raison d'être tient en une phrase : **le contenu de prévention existe, le format de quiz live existe — mais personne ne combine un corpus de prévention sourcé, une mesure par classe dans le temps, une éthique de minimisation des données (refus de toute surveillance intrusive), et une couche de détection complémentaire.** Notre pari n'est pas de concurrencer ; c'est d'assembler honnêtement et d'ajouter la couche manquante.

Gouvernance (bloc d'identité)

Champ	Valeur
Nom officiel	SOS ÉCRANS
Forme juridique	Association loi 1901, à but non lucratif
RNA	W751283606
SIREN	105 116 461 (statut actif)
Publication au JOAFE	Annonce n° 2041, parution n° 20260013 du **31 mars 2026**
Siège social	186 rue du Faubourg Saint-Martin, 75010 Paris
Contact public	contact@shieldy.org
Président	Emmanuel Klein
Trésorier	Largo Klein

L'association compte deux personnes — le président et le trésorier — et zéro salarié. Le projet est aujourd'hui porté par une **équipe associative très réduite** — ce qui explique un **développement progressif** et la **priorité donnée à la validation méthodologique avant tout déploiement large**. Nous le disons d'emblée, avec le concours d'outils d'assistance, et nous préférons que cette contrainte soit connue.

La méthode en trois lignes

1. **Traçabilité des sources.** Chaque affirmation pédagogique porte une source et un niveau de preuve ; sans source vérifiée, un contenu reste « brouillon » et n'est pas présenté à un élève.
2. **Validation humaine.** L'IA est un outil de production, pas une autorité. Elle prépare, l'humain tranche. Aucun contenu touchant au territoire clinique (la simulation sociale en particulier) n'est mis en classe sans validation préalable d'un pédopsychologue.
3. **Le témoin au centre, l'émotion jamais la terreur.** Toute la pédagogie est calibrée pour entraîner le geste du témoin, sans dramatisation ni culpabilisation.

Cinq faits-clés, sourcés

1. **Environ 18 %** des élèves déclarent avoir été victimes de cyberharcèlement, et **37 %** déclarent avoir subi du harcèlement et/ou du cyberharcèlement (données françaises 2024 ; INSEE Références 2025 / DEPP).
2. **25 % des 6–18 ans victimes** de harcèlement déclarent avoir déjà pensé au suicide ou à l'automutilation — **39 % chez les filles victimes** (e-Enfance / Caisse d'Épargne 2024). *À manier avec sobriété.*
3. **Un seul programme anti-harcèlement au monde dispose d'une preuve d'efficacité par essai contrôlé randomisé (RCT) : KiVa (Finlande)** — et il est payant, fermé, et sans quiz live (Kärnä et al., PubMed 23659182). Tous les autres dispositifs grand public sont des bibliothèques sans mesure d'efficacité.
4. **Environ 69 % des 12-17 ans** utilisent un téléphone Android (Arcep / CREDOC, *Baromètre du numérique 2024*) — proportion encore plus marquée dans les milieux modestes. C'est pourquoi notre effort de détection porte d'abord sur Android (où l'analyse du texte, sur l'appareil, est possible), pas sur iOS.
5. **Les plus exposés sont aussi les moins protégés** : le cyberharcèlement frappe sur les lignes de fracture existantes — filles, jeunes LGBTQI+, en situation de handicap, d'origine migrante, identités cumulées (risque ×2 à ×3) — et dans les foyers où la supervision et la littératie numérique parentales sont moindres (synthèse de recherche, voir §2 et §3).

Les manques, en une ligne

À ce jour : la détection grooming/manipulation est *codée* (et tourne sur l'appareil) mais non encore *validée* sur corpus français ; la chaîne d'alerte a été éprouvée sur un téléphone Android réel mais pas sur une diversité d'appareils, ni diffusée largement ; PREVENTO compte ≈ 20 questions entièrement sourcées sur 235 visées ; PREVENTO GO est un concept sans prototype ; aucun pédopsychologue n'a encore apposé son « Sceau » ; aucun pilote terrain n'a été mené ; le rescrit fiscal est en cours.

---

2. Le problème, chiffré et sourcé

Le cyberharcèlement n'est pas un sujet en quête d'existence : les États, les ONG et les grandes plateformes y consacrent des moyens. Les chiffres ci-dessous sont repris des fichiers de recherche du dossier ; nous indiquons la source à chaque fois et signalons ce qui est solide de ce qui l'est moins.

Prévalence (France, 2024).

- ≈ **18 %** des élèves déclarent avoir été victimes de cyberharcèlement ; **37 %** déclarent du harcèlement et/ou du cyberharcèlement (INSEE Références 2025, « Cyberviolences dans les établissements scolaires et dans la société » / DEPP-DGESCO).
- Concentration sur le secondaire (≈ 22 % collège, ≈ 29 % lycée), mais le **primaire monte vite (20 % en 2024 contre 13 % en 2023)** ; l'âge moyen des victimes baisse.

- **80 % des victimes connaissent et côtoient leur agresseur** — le cyberharcèlement n'est, dans la grande majorité des cas, pas le fait d'un inconnu. Nuance importante pour ne pas surinvestir la figure du prédateur extérieur.

Qui est touché — la donnée la plus robuste : les lignes de fracture.

- **Filles davantage exposées** aux cyberviolences à dimension sexuelle (rumeurs, images humiliantes) : **11 % des filles vs 7 % des garçons** (INSEE/DEPP).
- **Surexposition massive des jeunes LGBTQI+** (prévalence du cyberharcèlement de 10,5 % à 71,3 % selon les études), des jeunes en situation de handicap, d'origine migrante. **Effet d'intersection** : risque ×2 à ×3 chez les identités marginalisées cumulées (revues d'intersectionnalité NIH/PMC 2020–2025).
- **Niveau social — réel mais complexe.** Un enfant de 10 ans de milieu socio-économique faible aurait ~2× plus de risque de se déclarer cyberharcelé (EU Kids Online) ; mais la relation n'est pas linéaire et reste moins nette que pour le harcèlement physique. L'argument solide n'est pas un déterminisme social, mais le **déficit de protection** (supervision et littératie numérique parentales moindres). *Nous le signalons comme « à manier avec prudence ».*
- **Territoire (rural/urbain).** Pas de ventilation française fiable. Là où c'est mesuré (États-Unis), le rural est plutôt *plus* touché. *Nous ne l'affirmons pas pour la France.*

Gravité.

- Le lien cyber-victimation → idéation suicidaire / automutilation est confirmé par méta-analyses, *au-delà* du harcèlement traditionnel ; le **cumul** (scolaire + cyber) présente le risque de tentative le plus élevé (*European Child & Adolescent Psychiatry*, 2022).
- **25 % des 6–18 ans victimes** ont déjà pensé au suicide / à l'automutilation, **39 % chez les filles** (e-Enfance / Caisse d'Épargne 2024). Le suicide est la 2e cause de décès des 15–24 ans (Santé publique France 2023). *Nous citons ces chiffres sobrement et ne construisons aucun argumentaire sur la peur.*

Conséquence pour notre dispositif. Si les plus exposés sont aussi les moins protégés, alors l'outil doit aller vers eux : publics modestes, Android, relais écoles / REP / associations, et une détection qui ne repose pas sur la seule vigilance d'un parent déjà débordé.

---

3. Fondations scientifiques

Nous ne sommes pas psychologues. Nous nous appuyons sur des travaux publiés, et nous distinguons soigneusement ce qui est libellé exact / réutilisable de ce qui est seulement inspiration. Voici le socle.

3.1 Le harcèlement est un processus de groupe (Salmivalli)

Le cœur théorique de notre pédagogie est l'idée, fondée par Christina Salmivalli, que le harcèlement n'est pas un duo victime/agresseur mais un **phénomène de groupe** où chacun tient un rôle : meneur, assistant, renforçateur, témoin extérieur, défenseur.

- Source originale : **Salmivalli, Lagerspetz, Björkqvist, Österman & Kaukiainen (1996), *Bullying as a group process*, *Aggressive Behavior*, 22(1), 1–15** (paywall Wiley).
- Échelle réutilisable : la version auto-rapportée 15 items du **Participant Role Questionnaire** (attribuée à **Salmivalli & Voeten, 2004**) est reproduite *verbatim* dans un article en accès ouvert (PMC11851402) — ce qui la rend citable.

C'est de là que vient notre signature pédagogique : **placer le témoin au centre**, parce que c'est le groupe qui fait basculer une situation.

3.2 Le modèle du témoin (Latané & Darley)

L'entraînement du témoin s'appuie sur le **modèle séquentiel d'intervention du témoin de Latané & Darley** (les étapes : remarquer → percevoir l'urgence → se sentir responsable → savoir quoi faire → agir). C'est le cadre théorique explicitement repris par le programme FUSE (Irlande) et que nous adoptons pour structurer nos questions.

3.3 KiVa : le seul programme prouvé — et sa limite, dite honnêtement

KiVa (Université de Turku, Finlande) est le programme anti-harcèlement le plus étudié au monde, et le seul disposant d'une preuve d'efficacité par RCT.

- Résultats fondateurs (Finlande, 2007–2008, 234 écoles, ~8 000 élèves) : baisse de l'intimidation/victimation de **17–30 %** ; réplications en Italie (tailles d'effet 0,24–0,40) et aux Pays-Bas. Référence : **Kärnä et al. (PubMed 23659182)**.
- **Honnêteté requise : le RCT au Pays de Galles (2020) n'a PAS trouvé d'effet significatif** sur le critère principal (victimation auto-déclarée). L'efficacité dépend du contexte et de la fidélité d'implémentation. *Nous le citons tel quel — c'est exactement la posture que nous revendiquons.*
- Limite de transposition : KiVa est **payant et fermé** (le questionnaire opérationnel de Turku est sous licence ; ~20 £/élève la première année). Nous nous en inspirons sur l'**architecture** (la boucle de mesure annuelle baseline → post, restituée par classe), jamais sur le contenu.

Ce que nous retenons de KiVa pour PREVENTO : le rythme annuel baseline (T0) → post (T1), la double mesure élèves *et* enseignants (résultat *et* fidélité d'implémentation), et le « point de repère d'une année sur l'autre par classe » qui transforme un quiz en outil de pilotage.

3.4 Instruments de mesure (ce que l'on mesure, et avec quoi)

Notre choix méthodologique central : **mesurer l'auto-efficacité (« je me sens capable d'agir »), pas seulement la connaissance** — parce que l'auto-efficacité prédit mieux le passage à l'acte du témoin.

- **DABSS — Dublin Anti-Bullying Self-Efficacy Scales (Sargioti et al., 2022)**, sous-jacent au programme FUSE (Irlande). En **accès ouvert (PMC9969485)**. Structure : 4 échelles (victime/témoin × en ligne/hors ligne), chacune 20 items en 5 dimensions (reconnaître → comprendre l'urgence → se sentir responsable → savoir → intervenir), échelle de confiance 0–5, fiabilité $\alpha > .90$. C'est notre banque d'items à adapter (pas à copier) et le squelette du « profil radar à 5 branches » du tableau de bord.
- **rBVQ / OBVQ d'Olweus** (1996) — instrument de prévalence de référence (sous licence) : nous en reprenons la *logique de gabarit* (définir le harcèlement à l'élève avant de questionner ; cadre temporel fixe ; fréquences standardisées), pas le texte des items. Une adaptation française validée du rBVQ existe.
- **PRQ (Salmivalli & Voeten, 2004)** pour les rôles de groupe (cf. 3.1).
- **Digital Health Index belge (Safeonweb / CCB)** — modèle d'UX de restitution (score global + par thème + comparaison à une moyenne de référence + remédiation ciblée).

3.5 Corpus pédagogique réutilisable (licences vérifiées)

- **StopBullying.gov** (U.S. Dept. of Health & Human Services) — **domaine public** : définitions (les 3 critères : comportement agressif non désiré, déséquilibre de pouvoir, répétition ; les 3 spécificités du cyber : persistant, permanent, difficile à repérer).
- **Be Internet Awesome / Interland** (Google) — **CC BY 4.0** : la seule grande source mondiale clairement libre et réutilisable commercialement (vocabulaire enfant, leçons empathie, rôle du témoin).
- **eSafety** (Australie) — **CC BY 4.0** (« Rewrite Your Story » : réécrire la scène).
- **Kit ISC « Vivre ensemble »** (ANCT) — **Licence Etalab 2.0** : le principal contenu français libre.

- **Inspiration seulement (à ne pas copier)** : KiVa, Olweus, Common Sense (CC BY-NC-ND : ni traduire ni modifier), CEOP, Internet Matters, MediaSmarts (reproduction non lucrative mais pas d'adaptation), et les programmes coréen (Eoullim), japonais (kokoro / omoiyari).

3.6 Le cadre français : Catherine Blaya et Éric Debarbieux

Pour ancrer le dispositif dans la recherche française, nous nous appuyons sur les travaux de **Catherine Blaya** (cyberviolence, continuum offline/online, EU Kids Online France) et **Éric Debarbieux** (climat scolaire, facteur protecteur). Références identifiées dans le dossier : *La cyberviolence*, Que sais-je ?, PUF, 2025 (ISBN 9782715429345) ; « L'école à l'ère du 2.0 — Climat scolaire et cyberviolence » (PDF MEN, HAL halshs-03534707) ; et le **dataset CyberAgressionAdo-v1** (agressions en ligne annotées en français, HAL hal-03765860), que nous visons comme corpus de validation pour Shieldy (cf. §5).

Statut : Mme Blaya et M. Debarbieux n'ont pas été sollicités à ce jour. Ce sont des références que nous lisons et que nous souhaitons soumettre à leur regard critique — non comme une caution, mais pour identifier nos angles morts.

3.7 Fondations de la simulation sociale (PREVENTO GO)

PREVENTO GO mérite qu'on en expose la conception, parce que c'est la brique la plus singulière de l'écosystème — et la plus exposée au reproche de gadget. Ce n'en est pas un.

Une catégorie à part, revendiquée. PREVENTO GO n'est ni un jeu vidéo (pas de fun, pas de mascotte, pas de score), ni un quiz, ni un film (pas d'arc dramatique, pas de climax). Nous l'appelons **simulation sociale** : le joueur incarne un **témoin ordinaire** d'un groupe — jamais la victime, jamais le héros — et fait défiler un fil de messagerie sur cinq « jours ». Ce qu'on lui donne à voir n'est pas une agression spectaculaire, mais la **mécanique grise de l'exclusion** : un message lu sans réponse, un « mdr » de trop, un « vu à 23 h », des silences qui s'accumulent. Aucune insulte crue, aucun méchant désigné. À intervalles, une pause : *« qu'est-ce que tu fais ? »*, trois ou quatre choix dont **aucun n'est évidemment bon** (se taire protège socialement ; dire « arrêtez » coûte un risque). La fin est un **constat froid**, pas un score. **Le vrai produit, c'est le débrief** encadrant qui suit — l'objet-tiers (« Lucas », « Sami ») permettant de parler du système sans jamais désigner quiconque dans la classe.

Un moteur, pas une histoire figée. Au-dessus des épisodes, il y a un **moteur paramétrique** : des curseurs systémiques (vitesse de propagation, biais de la meute, présence d'alliés, visibilité de l'adulte, vitesse d'escalade) engendrent, à partir d'un même squelette, une **collection** de simulations adaptables à l'âge et au cadre. L'épisode pilote — **« Le Screenshot »** — est le premier d'une série possible, pas un objet isolé.

Fondé, pas improvisé. Le concept ne sort pas d'une intuition : il repose sur un **inventaire mondial des dispositifs de prévention** (KiVa, eSafety australien, Interland / Be Internet Awesome de Google, OK Groomer) et sur la **littérature de la gamification**, dont la convergence a fait émerger la catégorie « simulation sociale ». Il est encadré par une **doctrine de conception en cinq lois gravées** : (1) réalisme gris, jamais le spectaculaire ; (2) pas de punition morale — la neutralité doit offrir un bénéfice à court terme, le coût n'apparaissant que plus tard ; (3) la donnée est un miroir, pas un tribunal (« la pression du système a poussé vers la neutralité », jamais « 70 % ont été lâches ») ; (4) minimalisme austère, esthétique de notification, dans la lignée de *Papers Please*, *Reigns* ou *Her Story* — pas de Netflix ; (5) territoire clinique, donc **Sceau d'un pédopsychologue obligatoire** avant toute mise en classe.

L'honnêteté du statut. Le manifeste et l'épisode pilote sont **écrits** ; le **prototype reste à construire** (faux téléphone web, sans backend) et le guide de débrief — peut-être le document le plus important — à finaliser. Surtout : **aucune mise en classe ne se fera sans la validation préalable d'un pédopsychologue.** PREVENTO GO est un concept abouti et défendable, pas encore un objet jouable, et nous nous gardons de le présenter autrement.

4. Cadre éthique

Notre éthique n'est pas un supplément ; c'est la contrainte de conception. Six principes, gravés.

1. **Le Bouclier de vérité.** Chaque affirmation porte une source et un niveau de preuve (primaire / secondaire / tradition / reconstruction). Sans source vérifiée, un contenu reste « brouillon » et n'est jamais affiché à un élève. L'IA ne se valide jamais elle-même.
2. **Le Sceau humain.** Sur tout territoire clinique — la simulation sociale manipule du stress social, de l'empathie, de la cognition morale — **aucune mise en classe sans validation préalable d'un pédopsychologue ou d'un chercheur spécialiste du harcèlement.** L'IA prépare le dossier ; l'humain tranche. *Ce Sceau n'a pas encore été apposé : nous n'avons pas identifié de pédopsychologue à ce jour (cf. §7).*
3. **Émotion juste, jamais la terreur.** Pas de statistique anxiogène, pas de suicide brandi, pas de dramatisation. La séance finit toujours sur du constructif (la réparation), jamais sur la peur. Nous vendons la mémoire et l'entraide.
4. **Le témoin au centre — sans injonction héroïque.** La recherche dit que les témoins comptent ; nous ne laissons jamais entendre « si tu n'as rien dit, c'est ta faute ». Toute question sur le témoin s'accompagne du droit de ne pas agir seul, ni tout de suite.
5. **Empowerment plutôt que surveillance.** Côté Shieldy, refus explicite de toute intrusion disproportionnée dans la vie privée : *un enfant surveillé en cachette apprend à se cacher, pas à se protéger.* L'enfant sait que l'application est là (transparence absolue) ; l'alerte transmet le niveau et la nature du risque, jamais le texte ni l'identité d'un tiers (« contact inconnu » + catégorie).
6. **Le texte ne quitte pas le téléphone (Android).** Le texte des applications est analysé **sur l'appareil, sans envoi serveur** ; ce qui transite encore (l'analyse d'image et un assistant optionnel, en cours de retrait) est **analysé sans stockage durable**. Aucune donnée n'est vendue, jamais ; le **modèle économique ne repose ni sur la publicité comportementale ni sur l'exploitation commerciale des données des familles** (cf. §5.3 pour la vérité technique précise, par le code, et §6 pour les implications).

Garde-fous opérationnels complémentaires des quiz (déniabilité : l'élève répond sans s'auto-désigner victime ; *trauma-safe* ; chaque question sensible *skip_allowed* ; séquence respirable auditée dans son ensemble ; recours externe systématique au 3018 si le premier adulte n'aide pas) sont détaillés dans nos doctrines internes (`DOCTRINE-QUIZ-PREVENTO`, `MASTER-SIMULATION-SOCIALE`).

5. Fondations techniques (honnêtes)

5.1 Architecture mobile

> ***Aucune furtivité.*** *L'application est installée avec le parent,* ***visible en permanence*** *sur le téléphone,* ***activée volontairement****, et* ***l'enfant sait qu'elle est là*** *: c'est un contrôle parental* ***déclaré****, jamais caché.*

Le principe Shieldy : l'application **analyse localement** le **texte qui transite par le téléphone de l'enfant** — celui qu'il écrit, celui qu'il reçoit — et **détecte des signaux de risque** pour alerter le parent sur le *niveau* et la *nature* d'un risque. Elle ne casse aucun chiffrement, ne fait pas de surveillance réseau, et — c'est le point décisif sur Android — **n'envoie ce texte à aucun serveur.**

Android — c'est là que Shieldy donne sa pleine mesure

Sur Android, le système d'exploitation autorise ce que le bac à sable d'Apple interdit : qu'une application **repère des indices de risque**, avec le consentement explicite, parmi les signaux faibles circulant dans les autres applications. Shieldy s'appuie sur quatre piliers.

1. Le clavier Shieldy — capteur d'émission. Un clavier système (le clavier Shieldy) analyse le texte **au moment de la frappe, avant tout chiffrement et avant l'envoi**, dans *toutes* les applications qui l'utilisent : WhatsApp, Snap, Discord, SMS, jeux. C'est ce qui permet de couvrir les messageries chiffrées **sur les messages émis** — non en cassant le chiffrement (ce qui serait illégal et techniquement hors de portée), mais en lisant le texte *avant* qu'il parte. Le collage et la dictée vocale sont également captés. Un bouton SOS universel est intégré au clavier.

2. Le Service d'Accessibilité — capteur de réception. Là où le clavier ne voit que ce que l'enfant écrit, le Service d'Accessibilité **analyse localement ce qui s'affiche à l'écran** — donc les messages *reçus*. Il porte deux fonctions que le clavier ne peut pas assurer : la détection de **trajectoire** (une dérive qui se construit sur plusieurs messages reçus) et le **repérage d'un contact inconnu**, à partir des éléments affichés à l'écran propres à chaque application.

3. Le moteur de détection multi-couches (FR + EN), entièrement sur l'appareil. Le moteur d'analyse local couvre trois grandes familles : le **harcèlement** (insultes dirigées, menaces, *slurs*), le **grooming par phases** (modèle séquentiel d'O'Connell — mise en confiance → isolement → demande — décliné en 16 motifs tagués par phase), et la **sextorsion / le « brouteur »**. Sa règle de conception est centrale, et nous y tenons : **aucun signal isolé ne déclenche d'alerte.** Ce qui alerte, c'est la **trajectoire** — au moins trois phases observées, ou au moins deux assorties d'un contact inconnu. Les faux positifs sont activement écartés : cours de SVT, vocabulaire de jeu, paroles de chanson, *homoglyphes* (caractères visuellement identiques mais codés différemment). C'est la différence entre un filtre à mots-clés, qui inonde de fausses alertes, et un moteur de dynamique, qui attend de comprendre.

4. Le verrou anti-contournement. Un dispositif de protection qu'un enfant désactive en trois clics ne protège personne. Shieldy se défend à deux niveaux.

- *Sur tout téléphone (déployé)* : un **veilleur en arrière-plan** vérifie toutes les 15 minutes — et immédiatement au redémarrage du téléphone — que le clavier et l'accessibilité Shieldy sont toujours actifs. À la moindre coupure, il **prévient le parent** (SMS + e-mail), via une file d'attente persistante qui résiste à une perte de réseau. Le retrait des droits d'administration de l'appareil déclenche la même alerte.
- *Sur appareil dédié (cible institutionnelle)* : en mode **Device Owner** — un téléphone réinitialisé et provisionné pour l'enfant — la désinstallation et la désactivation deviennent **structurellement impossibles**. C'est le « vrai verrou », réservé à la distribution institutionnelle (par exemple les Cités éducatives) ; il reste un chantier de déploiement, distinct du watchdog qui, lui, fonctionne déjà partout.

L'alerte au parent transmet le **niveau** (orange / rouge) et la **nature** du risque (catégorie + « contact inconnu »), par SMS et e-mail — **jamais le contenu du message, jamais le nom d'un tiers.** Protéger sans dénoncer.

Ce n'est pas qu'une intention de code : la chaîne a été éprouvée en conditions réelles. Sur un téléphone Android physique (Redmi), un message de harcèlement saisi dans l'application a déclenché l'alerte et **l'e-mail au parent a effectivement été reçu** — chaîne validée en production (test du 22/05/2026). Un bug bloquant découvert à cette occasion (le serveur d'alerte rejetait le niveau en minuscules) a été corrigé dans la foulée.

Pourquoi Android, et pas d'abord iOS — un choix social autant que technique

Ce choix n'est pas un pis-aller, c'est une boussole. **Environ 69 % des 12-17 ans en France utilisent un téléphone Android** (Arcep / CREDOC, *Baromètre du numérique 2024*) — Apple ne domine que chez les 18-24 ans. Surtout, **Android est nettement surreprésenté dans les milieux les plus modestes**, où la supervision

parentale et la littératie numérique sont les plus difficiles à mobiliser (cohorte Elfe / Inserm ; travaux du Pr. Debarbieux ; en réseau d'éducation prioritaire, 4,4 % d'auteurs déclarés contre 2,5 % en moyenne nationale — DEPP, *Enquête harcèlement 2023*). Autrement dit : **les publics les plus exposés sont aussi les moins protégés, et ils sont sur Android.** Concentrer l'effort de détection sur Android n'est donc pas une facilité technique — c'est aller là où le besoin est le plus criant. Un outil qui ne fonctionnerait que sur iPhone tournerait le dos à ceux qui en ont le plus besoin.

iPhone — la limite, dite franchement

Apple a ses propres garde-fous parentaux (Temps d'écran, API FamilyControls), qui permettent de *bloquer ou filtrer* des applications. Nous avons développé pour iOS un navigateur sécurisé. Mais le bac à sable d'Apple **interdit hermétiquement à une application tierce de lire le contenu textuel circulant dans les autres applications** (iMessage, WhatsApp, Snap, Instagram). Une analyse sémantique du harcèlement y est donc **structurellement impossible** — non par retard de notre part, mais par architecture du système. **Sur iPhone, Shieldy ne couvre pas les applications natives ; nous le disons et nous y travaillons. C'est sur Android que Shieldy donne sa pleine mesure.**

5.2 Détection multi-couches — état réel

La détection est organisée en couches, du plus simple au plus complexe. Il est essentiel de distinguer ce qui est **déployé** de ce qui est **codé mais non validé**, et où chaque couche s'exécute (sur l'appareil ou sur serveur).

- **Déployé, sur l'appareil — moteur de motifs FR + EN.** Le cœur de la détection texte tourne **sur le téléphone**, dans le moteur d'analyse local, sans aucun appel réseau. Il couvre un large jeu de motifs (regex) : grooming explicite, manipulation émotionnelle (isolement, culpabilisation, *love bombing*, DARVO, gaslighting), désensibilisation sexuelle progressive, sextorsion, *flash grooming*, harcèlement entre pairs et *gaming messaging*. Ces motifs s'appuient sur des travaux publiés — **Street et al. (2024), arXiv 2409.07958** (détection de grooming par détermination de contexte et analyse au niveau message) ; le modèle des phases de grooming de **O'Connell** ; le **BF-PSR Framework** (Université de São Paulo) pour les signaux comportementaux (heure tardive, fréquence, questions intrusives). C'est la couche de motifs qui implémente la *règle d'or* : aucun signal seul n'alerte, seule la trajectoire le fait.
- **La limite honnête de cette couche.** Ces motifs sont **écrits à la main et n'ont pas encore été validés empiriquement sur un corpus français annoté.** La validation visée passe par le dataset **CyberAgressionAdo-v1** (Blaya et al., HAL hal-03765860) ; les premières mesures internes sur ce corpus donnent un rappel encore partiel (de l'ordre de la moitié des cas sur la couche pair-à-pair), ce qui confirme qu'**aucune couche unique ne suffit** et justifie l'empilement. Tant que cette validation systématique n'est pas faite, nous parlons d'un moteur *en développement*, pas *prouvé*.
- **Héritage serveur (Detoxify / Perspective).** Une génération antérieure de l'architecture faisait analyser le texte par un worker Cloudflare adossé à un modèle de toxicité multilingue (**Detoxify / multilingual-toxic-xlm-roberta**, Apache 2.0) et à **Perspective API (Google Jigsaw)**. Ce worker existe toujours dans le dépôt, mais **l'application Android ne lui envoie plus le texte des messageries** : l'analyse a été rapatriée sur l'appareil (cf. § 5.3). Cet héritage demeure une piste de renfort hors-ligne (modèle local embarqué), pas une dépendance d'exécution.
- **Annoncées, non intégrées en production.** L'adaptation française de séquences type *Roblox Sentinel* (Apache 2.0) et le fine-tuning sur le dataset *MentalManip* (ACL 2024) figurent dans notre feuille de route comme pistes ; ils ne sont pas en production validée.

5.3 Où le texte est analysé — la vérité, par le code

C'est l'affirmation la plus sensible d'un dispositif de protection de l'enfance, et nous la fondons sur l'inspection du code, pas sur une intention.

Sur Android, le texte des applications ne quitte pas le téléphone. Le clavier et le Service d'Accessibilité font appel au moteur d'analyse **localement** ; le code est explicite (« 100 % local — on-device — aucun envoi réseau »), et il n'existe dans l'application Android aucun appel au worker pour analyser le texte des messageries, ni aucun appel à HuggingFace. Le texte écrit ou reçu dans WhatsApp, Snap, Discord, les SMS ou les jeux est analysé sur place, puis n'est ni transmis ni conservé. C'est un argument **fort**, et il est vrai aujourd'hui.

Deux réserves, par souci d'exactitude :

- **Un assistant conversationnel optionnel (en cours de retrait).** Tant qu'il est présent, la requête que l'enfant lui adresse — par nature destinée à un service — transite par notre serveur, où une couche de sûreté l'analyse. Ce flux sort de l'appareil ; il est **analysé en mémoire, sans stockage durable**.
- **Les images.** La détection de contenu visuel (NSFW) sur les images reçues passe encore par un service (envoi de l'URL au worker `/analyze-image`). **C'est le dernier point où du contenu quitte l'appareil**, et c'est un chantier explicitement ouvert : faire passer cette détection *on-device* (ML Kit / TFLite) avant de pouvoir affirmer que *rien* ne quitte le téléphone. Nous ne le revendiquons donc pas encore pour les images.

Pour la part qui transite encore par nos serveurs (l'analyse d'image et un assistant optionnel, en cours de retrait), les principes valent sans exception : ces contenus sont transmis **sans aucun identifiant personnel** (ni nom, ni numéro de téléphone, ni compte), de façon **pseudonymisée**, **en transit et sans stockage** ; **aucune donnée n'est vendue** — le **modèle économique ne repose ni sur la publicité comportementale ni sur l'exploitation commerciale des données des familles** ; **aucun nom de tiers n'est jamais transmis** au parent (l'alerte dit « contact inconnu » et une catégorie). Les modèles tiers éventuellement mobilisés pour cette part résiduelle peuvent être opérés hors de l'Union européenne ; nous l'assumons et l'expliquons plutôt que de le dissimuler. Pour les fonctions d'**IA conversationnelle** (notamment l'aide à la préparation des séances PREVENTO par l'encadrant), nous faisons le **choix d'une solution française et souveraine — Mistral**. La trajectoire est claire : **le texte est déjà passé sur l'appareil ; les images suivront.**

5.4 Moyens, coûts et besoins

L'infrastructure technique est volontairement **frugale** : le site et les workers tournent sur Cloudflare (Pages + Workers), pour un coût marginal. C'est un choix d'économie assumé — mais ce n'est pas là que se situe le besoin.

Le coût réel, jusqu'ici, tient à deux choses : **le travail d'une personne seule sur près d'une année**, et **un investissement personnel déjà engagé** (plusieurs milliers d'euros en outils de conception et de développement). Le projet a été **entièrement autofinancé** à ce stade.

Pour passer d'un ensemble de prototypes à un dispositif **validé et déployable**, les besoins identifiés sont d'un autre ordre : accompagnement **juridique** (RGPD, analyse d'impact AIPD, cadre de la capture au clavier), **validation par un pédopsychologue**, **sourçage complet** du corpus pédagogique, et un **pilote de terrain mesuré**. C'est l'objet de la démarche de financement en cours.

---

6. Gouvernance et transparence

- **Structure.** Association loi 1901, **deux personnes** (président : Emmanuel Klein ; trésorier : Largo Klein), **zéro salarié**. Identité juridique vérifiable (RNA W751283606, SIREN 105 116 461, JOAFE du 31 mars 2026).

- **Données.** Sur Android, le texte des applications est **analysé sur l'appareil, sans envoi serveur** ; ce qui transite encore (l'analyse d'image et un assistant optionnel, en cours de retrait) est analysé **sans stockage durable**, transmis sans aucun identifiant personnel et de façon pseudonymisée. Aucune donnée vendue — le modèle économique ne repose ni sur la publicité comportementale ni sur l'exploitation commerciale des données des familles ; aucun nom de tiers transmis ; alerte limitée au niveau et à la nature du risque. PREVENTO est conçu sans compte, sans installation, anonyme.
- **RGPD.** Une analyse d'impact relative à la protection des données (AIPD) et une consultation juridique spécialisée — notamment sur l'articulation de la capture locale au clavier avec l'article 226-15 du Code pénal (secret des correspondances), et sur les droits des mineurs de 15 ans sur leurs propres données — sont **intégrées à notre feuille de route réglementaire, préalables à tout déploiement grand public**.
- **Fiscalité.** Le rescrit fiscal est **en cours**.
- **Documents officiels** (statuts, récépissé de préfecture, PV d'assemblée constitutive, liste des dirigeants, publication au JO) sont disponibles sur demande à contact@shieldy.org.
- **Modèle économique.** Gratuité absolue des outils, financement par dons et subventions. Cette gratuité est un choix éthique documenté (les inégalités face aux écrans sont attestées — INSERM/ELFE) autant qu'un positionnement.

---

7. Ce que nous ne savons pas encore / ce qui reste à valider

Cette section est, à nos yeux, la plus importante. Un dispositif sérieux se reconnaît à ce qu'il sait nommer ses limites.

1. **iOS partiellement aveugle — limite permanente.** Le bac à sable d'Apple interdit la lecture des messages des applications natives. Ce n'est pas un retard de développement : c'est structurel et durable. Sur iPhone, Shieldy ne couvre pas la détection de messages.
2. **Moteur de détection grooming/manipulation : codé, non encore validé sur corpus.** Les motifs (sur l'appareil) existent et la chaîne d'alerte est éprouvée, mais ils n'ont pas été validés systématiquement sur corpus français annoté. La validation sur CyberAgressionAdo-v1 (mesure réelle de précision/rappel par couche) reste à faire ; les premières mesures internes donnent un rappel encore partiel.
3. **Chaîne d'alerte validée sur un appareil ; reste à éprouver sur d'autres.** La chaîne complète (détection → alerte → e-mail parent reçu) a été validée en conditions réelles sur un téléphone Android physique (Redmi, 22/05/2026). Elle n'a pas encore été testée sur une diversité de constructeurs et de versions d'Android, ni diffusée largement. L'application iOS, elle, est compilée mais limitée par construction (cf. point 1).
4. **Angles morts assumés sur Android.** Nous les listons sans les masquer : (a) **audio** — les messages vocaux (WhatsApp, Snap) ne sont pas transcrits *on-device*, c'est un angle mort total ; (b) **images** — la détection NSFW passe encore par un service, son passage *on-device* est un chantier ouvert (tant qu'il n'est pas clos, nous ne disons pas que *rien* ne quitte le téléphone) ; (c) **applications hors-liste** — une messagerie comme Signal n'est pas encore couverte par les identifiants de vue ; (d) **désactivation** — l'accessibilité et le clavier restent désactivables sur un téléphone ordinaire (la coupure est *détectée et signalée au parent* par le watchdog, mais non *empêchée*) ; le verrou total (désinstallation impossible) n'existe qu'en mode Device Owner, sur appareil dédié.
5. **PREVENTO : ≈ 20 questions entièrement sourcées sur 235.** Le reste du corpus est en cours de sourçage sous Bouclier de vérité. Une question non sourcée reste « brouillon » et n'est pas présentée à un élève.
6. **PREVENTO GO : concept abouti, prototype à venir.** Le manifeste et l'épisode pilote (« Le Screenshot ») sont écrits ; il n'existe pas encore de prototype jouable ni de guide de débrief finalisé.

7. **Sceau pédopsychologue non apposé.** Aucun pédopsychologue ni chercheur spécialiste n'a, à ce jour, validé les contenus pour une mise en classe. La règle l'exige avant tout usage en classe de la simulation sociale.
8. **Aucun pilote terrain.** Le pilote (par ex. 10 classes réelles) — préalable à toute prétention d'efficacité — n'a pas été mené.
9. **Mesure de la compréhension, pas du comportement.** Nous mesurons l'évolution de la compréhension et de l'auto-efficacité par classe ; nous **ne mesurons pas le comportement réel**, et nous le disons. La preuve d'efficacité comportementale d'un programme se construit (modèle KiVa, étude coût-efficacité à 2 ans) ; elle ne se proclame pas.
10. **Cadre juridique RGPD / fiscal en cours.** AIPD, conseil pénaliste sur la capture clavier, et rescrit fiscal restent à finaliser.

---

8. Annexes

8.1 Bibliographie (références vérifiées, regroupées par thème)

> *Toutes les références ci-dessous proviennent des fichiers de travail du dossier et y figurent avec leur source. Les libellés exacts d'items ne sont réutilisés que lorsqu'ils sont publiés en accès ouvert.*

Cadre théorique — processus de groupe et rôle du témoin

- Salmivalli, C., Lagerspetz, K., Björkqvist, K., Österman, K. & Kaukiainen, A. (1996). *Bullying as a group process: Participant roles and their relations to social status within the group. Aggressive Behavior*, 22(1), 1–15. (Paywall Wiley.)
- Salmivalli, C. & Voeten, M. (2004). *Connections between attitudes, group norms, and behaviour in bullying situations. International Journal of Behavioral Development*, 28(3), 246–258. Items du PRQ (15) reproduits en accès ouvert : https://pmc.ncbi.nlm.nih.gov/articles/PMC11851402/
- Latané, B. & Darley, J. — modèle séquentiel d'intervention du témoin (cadre théorique repris par FUSE).

Mesure et efficacité

- Kärnä, A. et al. — essai contrôlé randomisé KiVa, effet sur la victimation y compris cyber : https://pubmed.ncbi.nlm.nih.gov/23659182/ ; programme : https://www.kivaprogram.net/
- RCT KiVa Pays de Galles (2020) — **absence d'effet significatif** sur le critère principal (à citer tel quel).
- Sargioti, A. et al. (2022). *Dublin Anti-Bullying Self-Efficacy Models and Scales* (DABSS), accès ouvert : https://pmc.ncbi.nlm.nih.gov/articles/PMC9969485/ ; programme FUSE : https://antibullyingcentre.ie/fuse/
- Olweus, D. (1996). *Revised Olweus Bully/Victim Questionnaire* (rBVQ/OBVQ) — instrument sous licence ; validation chilienne OBVQ-R : https://pmc.ncbi.nlm.nih.gov/articles/PMC8072054/
- Safeonweb / CCB — Belgian Digital Health Index : https://safeonweb.be/en/digital-health-index
- Méta-analyses d'efficacité des programmes : https://pmc.ncbi.nlm.nih.gov/articles/PMC8218972/

Corpus pédagogique (licences)

- StopBullying.gov (U.S. DHHS) — domaine public : https://www.stopbullying.gov/cyberbullying/what-is-it
- Be Internet Awesome / Interland (Google) — CC BY 4.0 : https://beinternetawesome.withgoogle.com/en_us/educators
- eSafety Commissioner (Australie) — CC BY 4.0 : https://www.esafety.gov.au/educators/classroom-resources

- Kit ISC « Vivre ensemble » (ANCT) — Licence Etalab 2.0 : https://lesbases.anct.gouv.fr/ressources/kit-atelier-favoriser-le-vivre-ensemble-et-prevenir-le-cyberharcelement-cycle-4-lycee-12-18-ans

Cyberharcèlement — France et Europe

- INSEE Références 2025, « Cyberviolences dans les établissements scolaires et dans la société » (SSMSI) ; DEPP-DGESCO.
- e-Enfance / Caisse d'Épargne 2024 (Audirep) ; Santé publique France 2023 ; Rapport Sénat harcèlement (2021).
- EU Kids Online & Net Children Go Mobile (LSE) ; JRC Commission européenne 2025 ; Görzig, Milosevic & Staksrud (2017) ; *European Child & Adolescent Psychiatry* (2022, méta-analyses) ; Tippett & Wolke (2014).
- Blaya, C. — *La cyberviolence* (Que sais-je ?, PUF, 2025, ISBN 9782715429345) ; « L'école à l'ère du 2.0 — Climat scolaire et cyberviolence » (PDF MEN, HAL halshs-03534707) ; dataset CyberAgressionAdo-v1 (HAL hal-03765860).
- Debarbieux, É. — « Du climat scolaire : définitions, effets et politiques publiques » (PDF MEN) ; « Refuser l'oppression quotidienne » (rapport MEN, 2011).

Détection technique

- Street, J. et al. (2024). *Enhanced Online Grooming Detection Employing Context Determination and Message-Level Analysis*, arXiv:2409.07958.
- O'Connell, R. — modèle des phases du grooming (notamment phase de désensibilisation sexuelle).
- BF-PSR Framework (Université de São Paulo) — signaux comportementaux de grooming.
- Detoxify / *unitary/multilingual-toxic-xlm-roberta* — modèle de toxicité multilingue, licence Apache 2.0 (HuggingFace).
- Perspective API (Google Jigsaw) — détection de toxicité en français.
- *Pistes non intégrées en production* : Roblox Sentinel (Apache 2.0) ; MentalManip (ACL 2024).

Cadre éducatif et programmatique (inspiration, non copiés)

- pHARe / « Non au Harcèlement » (éduscol) — cadre d'entrée dans les établissements.
- KiVa (Finlande), Olweus (Norvège), Eoullim (Corée), éducation morale / kokoro (Japon, NIER) — inspirations, non réutilisées telles quelles.

8.2 Glossaire éthique

- **Bouclier de vérité** — Règle selon laquelle toute affirmation porte une source et un niveau de preuve (primaire / secondaire / tradition / reconstruction). L'absence de source est affichée, jamais comblée. Un contenu non sourcé reste « brouillon » et n'est pas présenté à un élève.
- **Sceau (Sceau humain)** — Validation finale, par un humain compétent (pédopsychologue ou chercheur), obligatoire avant toute mise en classe d'un contenu touchant au territoire clinique. L'IA prépare le dossier ; elle ne s'auto-valide jamais. Le passage d'un statut incertain à « validé » est un acte humain conscient.
- **Témoin au centre** — Principe pédagogique déplaçant l'attention du duo victime/agresseur vers le groupe et le rôle du témoin (défenseur potentiel), parce que c'est le groupe qui fait basculer une situation — sans jamais transformer ce rôle en injonction héroïque ni en culpabilisation.
- **Émotion jamais la terreur** — On mobilise l'empathie et la mémoire, jamais la peur ; aucune statistique anxiogène, aucune dramatisation ; la séance finit sur du constructif.
- **Empowerment plutôt que surveillance** — L'enfant est acteur, informé et consentant, jamais objet d'une surveillance intrusive cachée ; l'alerte protège sans dénoncer (niveau + nature, jamais le texte ni l'identité d'un tiers).

- **Analyse on-device / non-stockage** — Sur Android, le texte des applications est analysé *sur l'appareil*, sans envoi serveur ; la part qui transite encore (l'analyse d'image et un assistant optionnel, en cours de retrait) est analysée sans stockage durable, sans aucun identifiant personnel et de façon pseudonymisée. Aucune donnée vendue — le modèle économique ne repose ni sur la publicité comportementale ni sur l'exploitation commerciale des données des familles ; recours assumé et explicité, pour cette part résiduelle, à des modèles tiers.
- **Auto-efficacité** — Croyance « je me sens capable de… » ; ce que nous mesurons en priorité, parce qu'elle prédit mieux le passage à l'acte que la connaissance factuelle.