PROYECTO OBJETO DE PARTICIPACIÓN

Premio AEPD a la protección de datos personales de colectivos vulnerables y frente a la violencia digital

Proyecto: TuyTu

Documento: Proyecto objeto de participación (B)

Promotor: Equipo promotor de TuyTu

Contacto: protegeme@tuytu.tech | https://tuytu.tech

Versión: 2.0

Fecha: 27/01/2026

Este documento describe el servicio, el flujo de uso, la arquitectura a alto nivel, la metodología preventiva, el tratamiento de datos, las medidas de seguridad, la gobernanza, la evaluación prevista y el plan de escalado, en coherencia con el objeto del premio.

Índice

1. Resumen del proyecto

TuyTu es un servicio de ciberseguridad orientado a la protección preventiva de imágenes personales frente a usos no consentidos por inteligencia artificial, con foco en la prevención de violencia digital asociada a deepfakes sexuales (por ejemplo, desnudos sintéticos o manipulación sexualizada).

El enfoque es sencillo: proteger antes de que el daño ocurra. Muchas medidas reactivas (denuncia, retirada, asistencia legal) son necesarias, pero llegan tarde. Yo trabajo en el punto anterior: aplicar una protección técnica a la imagen antes de su exposición o circulación para reducir su aprovechamiento por procesos automatizados de manipulación.

El servicio se usa desde el móvil. El usuario selecciona una imagen; la aplicación realiza una llamada a un servidor donde reside el algoritmo; el servidor devuelve la imagen protegida al dispositivo. No guardo imágenes de usuario: el tratamiento se limita al procesado y la devolución del resultado.

2. Adecuación al premio y objetivos

2.1 Adecuación al premio

Este proyecto encaja en el premio por tres razones: (1) protege datos personales (imágenes) en un contexto de alto riesgo, (2) se orienta a colectivos vulnerables frente a violencia digital, y (3) incorpora privacidad y seguridad desde el diseño.

2.2 Objetivos del proyecto

- Reducir el riesgo de uso no consentido de imágenes para manipulación sexualizada por IA.
- Elevar el coste del abuso y disminuir la eficacia de ciertos procesos automatizados de transformación basados en la imagen.
- Mantener un tratamiento de datos mínimo: procesar lo imprescindible y evitar almacenamiento de imágenes.
- Aportar un enfoque preventivo que complemente las vías de denuncia, retirada y apoyo a víctimas.

2.3 Alcance

- Protección de imágenes bajo solicitud del usuario (servicio on-demand).
- Procesamiento remoto en servidor con entrega del resultado al dispositivo.
- Medidas de seguridad y trazabilidad técnica minimizada (sin contenido personal).
- Mensajes claros de alcance, límites y buenas prácticas.

2.4 Qué no es

- No es una herramienta de retirada/denuncia automática en plataformas.
- No es un sistema de monitorización masiva de redes ni rastreo de contenido de terceros.
- No realiza reconocimiento facial ni identificación biométrica de personas.
- No promete invulnerabilidad: define límites y mitigaciones.

3. A quién protege: colectivos vulnerables y casos de uso

3.1 Colectivos vulnerables

- Menores y adolescentes (con enfoque de mediación parental y/o entornos educativos).
- Mujeres expuestas a violencia digital (acoso, difamación sexualizada, control coercitivo).

- Creadores de contenido y perfiles con alta exposición pública.
- Entornos educativos y comunitarios con necesidad de prevención (centros, asociaciones).

3.2 Casos de uso

Caso 1 — Menor/adolescente que comparte imágenes
- Riesgo: difusión no controlada y reutilización para humillación o manipulación sexualizada.
- Uso: proteger imágenes antes de publicarlas o compartirlas en grupos.
- Resultado esperado: reducir utilidad de la imagen para procesos automatizados de manipulación.

Caso 2 — Mujer con exposición pública o riesgo de acoso
- Riesgo: apropiación de fotos para deepfakes sexuales o campañas de difamación.
- Uso: proteger imágenes que se publican en perfiles o que circulan en contextos sensibles.
- Resultado esperado: elevar el coste del abuso y reducir la calidad/eficacia de transformaciones maliciosas.

Caso 3 — Creador/a de contenido
- Riesgo: scraping de imágenes y generación de material sexualizado no consentido.
- Uso: protección previa de imágenes antes de publicación.
- Resultado esperado: disminuir aprovechamiento automatizado por terceros.

Caso 4 — Centro educativo / asociación
- Riesgo: incidentes entre menores (difusión no consentida, ciberacoso, humillación).
- Uso: herramienta preventiva dentro de programas de convivencia y educación digital, con guía y protocolo.
- Resultado esperado: reducción de incidentes y mejora de conciencia preventiva.

4. Descripción del servicio (qué es y qué no es)

4.1 Propuesta de valor
- Prevención real: actúa antes del daño, no después.
- Uso simple: flujo corto desde el móvil, con un resultado inmediato.
- Privacidad por diseño: procesado bajo solicitud, sin almacenamiento de imágenes.
- Enfoque de impacto: prioriza colectivos vulnerables frente a violencia digital basada en imagen.

4.2 Funcionalidad principal
- Seleccionar imagen y obtener versión protegida.
- Guardar la imagen protegida en el dispositivo.
- Mensajes claros de alcance, límites y recomendaciones de uso.

4.3 Funcionalidades complementarias (roadmap)
- Protección por lotes para perfiles con gran volumen de imágenes.
- Perfiles de protección según escenario (familias, creadores, exposición pública).
- Material educativo y guías específicas para entornos educativos y familias.
- Integraciones futuras con flujos de almacenamiento/álbumes cuando haya viabilidad técnica y legal.

5. Flujo de usuario

5.1 Flujo principal (end-to-end)

18. Entrada: el usuario accede a la app y revisa información básica (qué hace, privacidad, límites).
19. Selección: el usuario elige una imagen del dispositivo.
20. Envío: la app envía la imagen al servidor por canal cifrado.
21. Procesado: el servidor aplica la protección y devuelve el resultado.
22. Entrega: el usuario guarda la imagen protegida en su dispositivo.
23. Uso: el usuario decide dónde la comparte o publica.

5.2 Puntos de control de privacidad

- Minimización: solo se envía la imagen necesaria para el procesado.
- No persistencia: la imagen no se conserva en el servidor tras el procesado.
- Logs minimizados: registros técnicos sin contenido de imagen.
- Transparencia: información clara para evitar expectativas irreales.

5.3 Comportamiento seguro en errores

- Si falla el procesado, el usuario puede cancelar y no se conserva la imagen en servidor.
- El servicio se diseña para evitar reintentos descontrolados o estados inconsistentes.
- Los errores se registran de forma técnica y minimizada para corregir fallos sin capturar contenido personal.

6. Arquitectura a alto nivel y flujo de datos

6.1 Componentes

- Aplicación móvil: interfaz, selección de imagen, envío y recepción de resultado.
- API segura: punto de entrada del servicio.
- Servicio de procesado: ejecución del algoritmo de protección.
- Observabilidad técnica: métricas agregadas y logs minimizados.
- Controles de seguridad: control de acceso, segmentación y hardening.

6.2 Flujo de datos (alto nivel)

Móvil (imagen seleccionada) → API segura → Servicio de procesado → Respuesta (imagen protegida) → Móvil.
En paralelo, se generan registros técnicos de funcionamiento sin incluir contenido de imagen.

6.3 Decisiones de diseño relevantes para privacidad

- No se crea un repositorio de imágenes: evita concentrar datos sensibles.
- El tratamiento está limitado a una finalidad única: devolver la imagen protegida.
- Se prioriza infraestructura con garantías RGPD y controles de acceso estrictos.

7. Metodología preventiva de protección de imágenes (visión conceptual)

7.1 Modelo de amenaza

- Apropiación de imágenes (públicas o privadas) y uso para generar deepfakes sexuales.
- Scraping y procesado automatizado de imágenes de perfiles públicos.

- Reutilización para suplantación o humillación.

7.2 Principio de actuación

La protección se basa en aplicar una transformación controlada sobre la imagen para conservar su utilidad visual para una persona, pero interferir en patrones explotables por determinados procesos automatizados de IA. El objetivo es reducir la eficacia de la generación o transformación maliciosa basada en esa imagen, elevando el coste del abuso.

7.3 Robustez y evolución

- El ecosistema de IA evoluciona rápido: el servicio se plantea como un sistema vivo, con pruebas periódicas y ajustes.
- Se contempla robustez frente a compresión, recortes y transformaciones comunes, porque son habituales en redes sociales.

8. Tratamiento de datos personales y privacidad desde el diseño

8.1 Datos tratados

- Imagen enviada por el usuario para obtener una versión protegida.
- Metadatos técnicos mínimos necesarios para operar y diagnosticar fallos (sin incluir contenido de imagen).

8.2 Datos no tratados / no almacenados

- No se almacenan imágenes de usuario en servidor tras el procesado.
- No se construyen perfiles ni repositorios de imágenes.

8.3 Finalidad y minimización

La finalidad es única: aplicar la protección preventiva a la imagen solicitada por el usuario. La minimización se aplica evitando cualquier tratamiento adicional no necesario, especialmente la conservación de imágenes.

8.4 Conservación

- Imágenes: procesado y descarte.
- Registros técnicos: retención limitada y acceso restringido, orientados a seguridad y fiabilidad del servicio.

8.5 Transparencia y derechos

- Información de privacidad en lenguaje claro, especialmente en escenarios de vulnerabilidad.
- Canal de contacto para ejercicio de derechos.
- Procedimiento interno para responder solicitudes y documentarlas.

8.6 Evaluación de impacto

Por tratar imágenes en un contexto de violencia digital y con posible enfoque a menores, se contempla una evaluación de impacto (DPIA) antes de despliegues amplios o pilotos que lo requieran, para reforzar cumplimiento y seguridad.

9. Medidas de seguridad (técnicas y organizativas)

9.1 Medidas técnicas

- Cifrado en tránsito (HTTPS/TLS).
- Eliminación tras procesado: sin persistencia de imágenes en servidor.
- Control de accesos al entorno de procesado bajo mínimo privilegio.
- Separación de entornos (desarrollo, pruebas y producción).
- Registro técnico minimizado sin contenido personal.
- Hardening del servidor y gestión de vulnerabilidades.

9.2 Medidas organizativas

- Política interna de accesos y revisiones periódicas.
- Gestión de incidencias y continuidad del servicio.
- Evaluación de proveedores con garantías RGPD.

9.3 Medidas orientadas a colectivos vulnerables

- Mensajes de advertencia y alcance claros, para evitar falsas expectativas.
- Configuración por defecto conservadora.
- Materiales de acompañamiento para familias y centros educativos.

10. Gobernanza, transparencia y gestión de incidencias

10.1 Gobernanza

- Revisión periódica de riesgos y controles, con registro interno de decisiones.
- Control de cambios: cambios relevantes en tratamiento o arquitectura se documentan.
- Responsabilidades internas de privacidad y seguridad definidas dentro del equipo.

10.2 Gestión de incidencias

- Procedimiento para detectar, contener y resolver incidencias técnicas y de seguridad.
- Evaluación de impacto y, cuando aplique, cumplimiento de obligaciones de notificación.
- Comunicación transparente a usuarios si existiera riesgo para sus datos.

10.3 Comunicación responsable

- No se promete invulnerabilidad; se explican límites y recomendaciones complementarias.
- Prevención combinada: técnica, educativa y reactiva.

11. Evaluación prevista: técnica, usabilidad e impacto social

11.1 Evaluación técnica

- Pruebas internas con conjuntos de imágenes controlados y con consentimiento.
- Pruebas ante compresión y transformaciones habituales en plataformas.
- Validación periódica para medir degradación de resultados de manipulación, en entorno controlado.

11.2 Usabilidad y comprensión

- Medición de facilidad de uso: tiempo de completado, errores, fricción.
- Medición de comprensión de alcance y límites: evitar malentendidos.

- Feedback cualitativo con perfiles objetivo (familias, creadores, entornos educativos).

11.3 Impacto social (cualitativo)
- Indicadores de confianza y sensación de control percibido por usuarios.
- Adopción y repetición de uso como señal de utilidad preventiva.
- Feedback sobre utilidad dentro de protocolos de prevención en centros/organizaciones.

12. Originalidad e innovación
- Prevención aplicada a imagen: intervención antes del daño, no solo reacción.
- Privacidad por diseño operativa: tratamiento mínimo y sin almacenamiento de imágenes.
- Enfoque directo sobre violencia digital sexualizada, con uso accesible desde móvil.
- Modelo complementario: combina prevención técnica con educación y vías reactivas.

13. Cómo reduce el daño en colectivos vulnerables

13.1 Punto de intervención en la cadena de daño
La cadena típica de daño comienza con la disponibilidad de imágenes, su recolección y el procesado automatizado para generar material manipulativo. TuyTu interviene al inicio, reduciendo el valor de la imagen como insumo para manipulación automatizada y aumentando el coste del abuso.

13.2 Beneficios esperados (realistas)
- Menos probabilidad de obtener manipulaciones eficaces y de calidad sobre imágenes protegidas.
- Menor exposición a sextorsión cuando el atacante no logra resultados de calidad.
- Refuerzo de prevención en familias y centros, combinando herramienta y educación digital.

14. Plan de escalado y sostenibilidad operativa

14.1 Despliegue por fases
24. Fase 1: MVP estable con flujo principal y controles de privacidad/seguridad consolidados.
25. Fase 2: mejora de robustez, UX y capacidades complementarias (por ejemplo, lotes y perfiles).
26. Fase 3: integraciones con servicios de almacenamiento y flujos de uso cuando haya garantías técnicas y legales.
27. Fase 4: despliegue de materiales y programas preventivos para entornos educativos y comunitarios.

14.2 Sostenibilidad operativa
- Modelo freemium: acceso básico y funciones avanzadas para perfiles de mayor necesidad de protección.
- Colaboración con ecosistemas de prevención y concienciación digital.
- Confianza como base: privacidad, seguridad y transparencia.

15. Riesgos, límites y mitigaciones

15.1 Límites técnicos
- No existe defensa universal frente a todos los métodos adversarios.

- Recaptura (foto a pantalla) y transformaciones extremas pueden reducir la protección.
- La eficacia puede variar según el modelo/ataque; por eso se plantea evaluación continua.

15.2 Riesgos de privacidad

- Errores de configuración: mitigación con cifrado, revisión de cambios y pruebas.
- Exposición accidental por logs: mitigación con minimización y controles estrictos de acceso.

15.3 Riesgos de comunicación y uso indebido

- Falsa sensación de seguridad: mitigación con mensajes claros y recomendaciones complementarias.
- Abuso del servicio por automatización: mitigación con limitación de tasa y monitorización técnica.