Cegados por la IA

Anexo D: Matriz de Riesgos y Filosofía Ética

Índice

1. Tabla de Riesgos Críticos 1
2. Riesgos a gran escala 3
3. Plausibilidad: Partes Frágiles del Proyecto 4
4. Líneas Rojas (Prohibiciones Estrictas) 5
5. Paradas de Emergencia 5
6. Rendición de Cuentas 5
7. Postura Ética del Equipo 6
8. Límites de la Personalización (Hard Constraints) 6
9. Referencias 7

1. Tabla de Riesgos Críticos

Hemos sido "abogados del diablo" con nuestro propio proyecto. Estos son los riesgos reales y cómo los mitigamos.

Nombre del Riesgo	Descripción	Gravedad (1-5)	Probabilidad (1-5)	Acciones de Prevención	Acciones de Contingencia
1. Violación de Privacidad de Terceros	Grabar y procesar imágenes de personas en el espacio público o en zonas sensibles (baños, vestuarios) sin su consentimiento.	5 Crítica	5 Constante	**Procesamiento 100% local (Edge AI)**. No hay almacenamiento de imágenes. Difuminado de caras de terceros por defecto. Detección y desactivación automática en zonas sensibles (mapeo de baños, etc.).	El sistema no guarda datos, por lo que no hay "fuga" que gestionar. Si se reporta un fallo en la desactivación, se priorizará una **actualización de software** para corregir la detección de zonas.
2. Descripciones Sesgadas o Estigmatizantes	Que la IA haga **descripciones subjetivas u ofensivas** basadas en estereotipos ("parece	5 Crítica	3 Media	Entrenamiento del modelo enfocado en la **objetividad descriptiva**. Fase de pruebas intensiva con	**Canal de reporte** inmediato (vía app o voz) para "descripción ofensiva". Revisión humana del *reporte* (no de la imagen) y

	sospechoso", "viste mal").			la ONCE y voluntarios para identificar y corregir sesgos antes del lanzamiento.	actualización prioritaria del modelo local.
3. Fallo Crítico de Seguridad (Error)	**Identificación incorrecta de un peligro** inminente (ej. no ver un coche, confundir un escalón).	5 (Crítica)	2 (Baja)	Formación inicial (con la ONCE) al usuario. Es una **ayuda complementaria** y no sustituye al bastón o perro guía. Priorizar la fiabilidad en objetos clave (obstáculos,etc).	Si un usuario reporta un fallo de seguridad, se investigará el contexto reportado y se lanzará una **actualización de emergencia** del modelo de detección.
4. Dependencia Tecnológica Excesiva	Que el usuario abandone el uso del bastón, perro guía o la lectura en Braille, generando una dependencia total del sistema.	3 (Media)	4 (Alta)	**Programa de formación** que fomenta el uso como complemento. El sistema recordará activamente la importancia de usar otras ayudas. No ocultar la posibilidad de fallo.	**Estudios de seguimiento** post-lanzamiento para evaluar la dependencia. Si se detecta, reforzar los programas de formación y concienciación con la ONCE.
5. Viabilidad Técnica en 2029	Que la **suposición del procesamiento local (Edge AI) falle**: el chip se sobrecalienta, la batería no dura, o no es lo bastante potente para el modelo.	4 (Alta)	3 (Media)	Diseño de modelos eficientes (quantized). Investigación activa sobre el hardware de 2029 (ej. sucesores de Gemini Nano). Prototipado constante.	Tener un "Plan B" con un modelo más ligero y con menos funciones (ej. sólo detección de obstáculos) que sí pueda correr en local, sacrificando la descripción detallada.
6. Venta de la Empresa y Mal Uso	Que la **empresa sea comprada** (ej. por Zara) y el nuevo dueño intente meter publicidad ("Ese polo es de Zara") o cambiar la política de privacidad.	4 (Alta)	2 (Baja)	**Licencia de software restrictiva**. Acuerdos contractuales con los usuarios y la ONCE que blinden la política de "no datos" y "no publicidad".	**Acciones legales** basadas en la licencia. **Transparencia total** con los usuarios sobre el cambio de propiedad y defensa de sus derechos adquiridos.
7. Obsolescencia del Modelo Local	El **modelo** se queda **anticuado** (aparecen nuevos objetos) y no podemos reentrenarlo eficazmente porque no recolectamos datos de los usuarios.	3 (Media)	4 (Alta)	El reentrenamiento se basará en **reportes de fallos** de los usuarios (ej. "no identifica las nuevas bicicletas públicas") y en datos sintéticos.	Si los reportes indican que el modelo está obsoleto, la empresa se compromete a lanzar **actualizaciones periódicas** (ej. anuales) del modelo base, entrenadas "en casa" (no con datos de usuarios).

8. Inaccesibilidad por Coste	Que las gafas sean tecnológicamente muy caras y solo accesibles para una élite, creando una nueva brecha de desigualdad.	3 (Media)	4 (Alta)	Modelo de negocio basado en **acuerdos con organizaciones** (como la ONCE) y sistemas públicos de salud para subvencionar o cubrir el coste del dispositivo.	**Buscar acuerdos** de financiación o planes de "renting" social si los acuerdos iniciales no son suficientes para cubrir la demanda de personas sin recursos.

2. Riesgos a gran escala

Vamos a exponer algunos riesgos posibles que potencialmente podría tener nuestro sistema si fuera usado de forma masiva en todo el mundo.

1. **Errores sistemáticos de percepción (millones de descripciones incorrectas)**

Incluso un 1% de error (que sería un resultado bueno para visión artificial) implicaría millones de descripciones mal generadas cada día.

Problemas que puede generar:

- **Confundir objetos** peligrosos, p. ej., no detectar un coche o describirlo demasiado tarde.
- **Describir mal** señales, precios, medicamentos, alimentos, causando daño en la vida cotidiana.
- **Sesgos** sistemáticos: reconocimiento peor de rostros, objetos o entornos propios de ciertas culturas o regiones.
- **Falsas sensaciones de confianza**, llevando a la persona usuaria a arriesgarse más de lo que debería.

Colectivos afectados:

- Personas con discapacidad visual: impacto directo en seguridad y autonomía.
- Aseguradoras y sistemas sanitarios: aumento de accidentes en usuarios.

2. **Dependencia excesiva del sistema**

Si cientos de millones de personas dependen de Cegados por la IA, un fallo global podría paralizar a toda una población.

Problemas que puede generar:

- **Caídas del servidor**: millones de personas pierden su “percepción asistida”.
- **Actualización defectuosa**: “ceguera digital” simultánea.
- **Ataques** que incapaciten el servicio.

3. **Riesgos de privacidad masivos**

Aunque solo capture vídeo, la escala multiplica el problema: millones de cámaras andando por el mundo, grabando espacios públicos y privados.

Problemas que puede generar:

- **Exposición accidental** de escenas íntimas de terceros.
- **Reconocimiento involuntario** de personas: riesgo de trazabilidad masiva.
- Datos extremadamente sensibles si alguien consigue acceso (hacking, **filtraciones**).

Colectivos afectados:

- Ciudadanía general, incluso quienes no usan el producto.
- Negocios, escuelas, hospitales: espacios donde la grabación constante es un riesgo.

4. **Problemas legales y regulatorios globales**

Hasta ahora hemos especificado nuestro sistema teniendo en cuenta las leyes españolas y europeas, pero si se expande su uso a nivel mundial puede haber otras leyes que no hemos tenido en cuenta.

Problemas que puede generar:

- **Legislaciones diferentes** (UE, EE.UU., India, etc.) entrarían en choque.
- Obligación de borrar datos, exigir consentimiento, informar a terceros de grabación…

3. Plausibilidad: Partes Frágiles del Proyecto

Aquí definimos los principales riesgos, su impacto, probabilidad y las acciones que tomaremos. Lo más dudoso del proyecto se basa en nuestra apuesta tecnológica a 2029:

1. **El Procesamiento 100% Local (Edge AI)**: Esta es la piedra angular de nuestra garantía de privacidad. Dependemos totalmente de que en 2029 los chips móviles sean capaces de ejecutar modelos multimodales complejos en tiempo real, sin drenar la batería en 30 minutos y sin sobrecalentarse. Hoy por hoy, esto es inviable. Es una apuesta tecnológica fuerte.

2. **La Mejora del Modelo "a ciegas"**: Decimos que no recolectamos datos, pero que mejoraremos el modelo con "reportes de usuarios". Esto es algo frágil. ¿Cómo corregimos un sesgo o un error de identificación si no podemos ver la imagen que lo causó? Confiar solo en la descripción verbal del usuario para un fallo visual es un ciclo de vida de producto muy dudoso y lento.

3. **La Detección de Zonas Sensibles:** Asumir que las gafas "sabrán" que están en un baño o un vestuario para desactivarse solas es un desafío de IA en sí mismo. Requiere un reconocimiento de escenas casi perfecto, y un fallo aquí supone un riesgo de privacidad.

4. Líneas Rojas (Prohibiciones Estrictas)

Estos son los límites que "Cegados por la IA" nunca cruzará:

1. **NUNCA se almacenará vídeo o imagen.** Ni en local ni en la nube. El procesamiento es 100% efímero (en tiempo real) y los datos se destruyen al instante.
2. **NUNCA se enviarán datos a la nube para procesar.** Todo el análisis de IA ocurre dentro de las gafas del usuario.
3. **NUNCA se incluirá publicidad.** El sistema es una herramienta de asistencia, no una plataforma publicitaria. La confianza del usuario es prioritaria.
4. **NUNCA se identificarán caras.** El sistema podrá decir "hay una persona", pero no "es Juan Pérez". El difuminado de caras de terceros será una prioridad técnica para proteger a los transeúntes.
5. **NUNCA se tomarán decisiones por el usuario.** Las gafas describen ("coche acercándose"), no ordenan ("cruza ahora" o "muévete"). La autonomía y responsabilidad final es siempre del usuario.

5. Paradas de Emergencia

Es vital que el usuario tenga control total sobre el sistema.

- **¿Quién puede parar el sistema?** Únicamente el usuario portador de las gafas.
- **¿Cuándo (Condiciones)?** En cualquier momento y por cualquier motivo (ej. privacidad, petición de un tercero, información irrelevante o abrumadora, entrada a un domicilio privado).
- **¿Cómo (Procedimiento)?** Apagado de IA (Total): Una pulsación larga (3 segundos) del mismo botón. El sistema de IA se apaga completamente. Las gafas quedan "dormidas" y no procesan ninguna imagen hasta que el usuario las reactiva.

6. Rendición de Cuentas

- **¿Quién rinde cuentas?** La empresa desarrolladora de "Cegados por la IA".
- **Canal de Reclamación:**
    - *Qué:* Reportar descripciones erróneas, sesgadas u ofensivas; fallos de seguridad (no detectar un obstáculo); o cualquier queja sobre privacidad o funcionamiento.

    - *Cómo:* Mediante un canal telefónico accesible (gestionado en colaboración con la ONCE) y a través de una función de "Reportar fallo" en la app móvil vinculada.
    - *Plazos:* Acuse de recibo en 24h. Evaluación inicial y respuesta en 72h.
- **Revisión Humana:**
    - Garantizamos la revisión humana de *todas* las reclamaciones y reportes de fallos. Como no tenemos los datos de imagen, la revisión se basará en el reporte del usuario y en los logs anonimizados del sistema (ej. "el modelo tuvo un 55% de confianza al identificar 'coche' en el momento del reporte").
- **Formas de Reparación:**
    - *Reversión:* No aplica (no hay datos que borrar).
    - *Corrección:* Si se confirma un fallo del modelo (error o sesgo), se prioriza su corrección y se distribuye una actualización de software a todas las gafas.
    - *Compensación:* Si un fallo del sistema causa un daño físico o material demostrable, se activará el seguro de responsabilidad civil de la empresa para compensar al usuario afectado.
    - *Prevención:* Las correcciones aplicadas al modelo sirven como prevención para que el fallo no vuelva a ocurrirle a ningún usuario.

7. Postura Ética del Equipo

Postura de cada miembro del grupo:

- Diego Alonso Arceiz - Team Advocate: **safetyist**
- Carmen Fernández González - Product Owner: **skeptic**
- Ignacio Gutiérrez Sánchez - Scrum Master: **skeptic**
- Bryan Xavier Quilumba Farinango - Experto Git: **accelerationist**

En el debate interno entre *Accelerationism* (innovar a toda costa) y *Safetyism* (parar por miedo), adoptamos una **postura intermedia y pragmática**.

- No frenamos la tecnología porque los beneficios de autonomía para un ciego son inmensos.
- Pero imponemos "cinturones de seguridad": auditorías externas y rechazo a la personalización ideológica.

8. Límites de la Personalización (Hard Constraints)

En Cegados por la IA una personalización excesiva podría llevar a:

- Interpretar la realidad de forma sesgada.
- Reforzar valores problemáticos (prejuicios).
- Aumentar los riesgos de dependencia o manipulación.
- Crear desigualdad entre usuarios con outputs radicalmente diferentes.

Si se permite **personalización profunda**, el sistema corre el **riesgo de convertirse en un filtro ideológico y comportamental** de la realidad, amplificando sesgos, dependencias y riesgos sociales. Si se **limita** demasiado, puede ser **percibido como rígido o poco útil**.

El equilibrio razonable para “Cegados por la IA” es:

- Personalización funcional → SÍ.
- Personalización interpretativa o ideológica → NO.

La personalización de nuestro sistema se hará a través de la **app móvil**, por medio de un menú de ajustes que estará definido por nosotros, por tanto las opciones de personalización estarán limitadas.

Límite 1: Personalización solo superficial

- El sistema NUNCA adapta la interpretación de la realidad.
- Solo adapta:
    - Velocidad
    - Nivel de detalle
    - Orden de la información
    - Preferencias contextuales

Límite 2: Prohibir personalización ideológica, moral, identitaria o discriminatoria.

- El usuario no puede pedir:
    - Describir a las personas de manera distinta según rasgos sensibles
    - Ocultar información que pueda ser relevante para su seguridad
    - Enfatizar prejuicios o valoraciones subjetivas

Límite 3: Modelos que no guardan historial sensible

- No se almacena:
    - Comportamiento en espacios públicos
    - Personas que encuentra habitualmente
    - Rutas de desplazamiento
    - Objetos que posee o usa

Límite 4: El modelo no hace juicios de valor

- Nunca debe decir cosas como:
    - “Esa persona parece peligrosa / poco confiable”
    - “Este lugar parece de mala categoría”
    - “Esa persona parece triste / nerviosa / agresiva”

Límite 5: Política explícita de “seguridad del tercero no usuario”

- La personalización nunca puede perjudicar a terceros grabados involuntariamente.

9. Referencias

- **Documentos de los sprints:**
    - *Análisis de Riesgos y Límites v2.0.docx*

- *Desafíos éticos v3.0.docx*

- **Referencias externas:**
  - Gestión de Riesgos en IA (IBM): https://www.ibm.com/es-es/think/insights/ai-risk-management
  - Artículo 9 del AI Act (Sistemas de alto riesgo): https://artificialintelligenceact.eu/es/article/9/
  - The Age of AI Anxiety (Resistance): https://journalofdemocracy.org/articles/the-age-of-ai-anxiety/
  - Conversación IA - https://chatgpt.com/share/691b5e80-cad4-8002-a235-abd0b71905f7
  - AI Triad: A Dialogue Across Differences
  - The Age of AI Anxiety — and the Hope of Democratic Resistance