Cegados por la IA - Model card

1. Resumen del modelo y del proyecto

Nombre del modelo: Cegados por la IA - **Versión:** 1.0 - **Fecha:** 08/12/2029
Autores: Carmen Fernández (Product Owner), Diego Alonso (Team Advocate), Ignacio Gutiérrez (Scrum Master), Bryan Xavier Quilumba (Experto Git).

Cegados por la IA es un sistema integrado en **gafas inteligentes** que transforma vídeo en descripciones auditivas en tiempo real para usuarios con discapacidad visual.

Representamos a la **empresa desarrolladora**, cuyo objetivo es facilitar la accesibilidad a personas con discapacidad visual, abordando una necesidad social persistente: la falta de una interacción más segura, autónoma y fluida en espacios públicos y privados.

Esta idea surge de la **demanda creciente de apoyo tecnológico a personas discapacitadas**, apoyada en la evolución de dispositivos portables (*on-edge*). A diferencia de apps móviles o métodos tradicionales, ofrece percepción continua, manos libres, priorización de eventos críticos y robustez frente a ruido, así como protección reforzada para datos de terceros captados por las gafas.

El modelo combina **visión por computador** para la detección de objetos, **LLMs** para la generación de descripciones y **TTS** (*Text-to-Speech*) para la síntesis de voz en tiempo real.

2. Casos de uso previstos y no previstos

El sistema está diseñado para asistir a personas con discapacidad visual o visión reducida.

Previstos

- **Navegación en exteriores e interiores:** Detección de obstáculos, pasos de cebra…
- **Lectura de textos:** Reconocimiento de precios, menús, documentos y pantallas.
- **Interacción social:** Descripción general del entorno y presencia de personas.

No previstos

- **Vigilancia masiva:** No se permite grabación continua ni almacenamiento de vídeo.
- **Identificación biométrica no consentida:** No se permite el reconocimiento facial de terceros desconocidos para evitar riesgos.
- **Sustitución de seguridad:** No debe usarse como única herramienta de seguridad en situaciones de riesgo vital sin ayudas complementarias.
- **Personalización ideológica:** Se prohíbe configurar el modelo para emitir juicios de valor moral, político o discriminatorio sobre el entorno.

3. Datos y entrenamiento

El entrenamiento se ha realizado utilizando **datasets públicos** diversos, como escenas urbanas y vídeos en primera persona, garantizando diversidad de entornos y culturas.

- **Privacidad en el entrenamiento:** El sistema **NO utiliza datos de los usuarios finales** para reentrenar el modelo de forma directa.
- **Ciclo de mejora (*Feedback Loop*):** La mejora del modelo se basa exclusivamente en **reportes voluntarios de texto** enviados por usuarios. Estos reportes se convierten en **datos sintéticos** en entornos controlados para el reentrenamiento.

Para garantizar la viabilidad en un dispositivo *edge*, partimos de un modelo multimodal preentrenado al que se aplica *fine-tuning*.

- **Arquitectura:** Modelo compacto de aproximadamente **1.3B de parámetros**, para equilibrar fluidez de lenguaje y capacidad de ejecución local.
- **Cómputo estimado:** El proceso de entrenamiento completo (incluyendo iteraciones y pruebas) se estima en unas **624 horas de GPU A100**.
- **Coste económico:** Estimado entre 2.000 y 3.000 USD (considerando costes de computación en la nube y almacenamiento).

Durante el entrenamiento, se estima una emisión de **~100 kg de CO_2e**, derivada de un consumo energético aproximado de **400 kWh.** Las gafas usan chips neuromórficos de bajo consumo, minimizando la huella de carbono operativa comparada con soluciones basadas en la nube.

4. Evaluación y métricas

El sistema utiliza un umbral de confianza estricto para la generación de descripciones:

- **Umbral de seguridad:** Si la probabilidad de acierto en la detección de un objeto es baja, el sistema emite una descripción genérica ("objeto desconocido") o verbaliza su incertidumbre ("no estoy seguro").
- **Alucinaciones:** Se penaliza severamente la "invención" de información en la función de pérdida durante el entrenamiento del LLM.

Antes de cualquier actualización del firmware:

- **Pruebas con usuarios reales:** Validación en colaboración con la ONCE para asegurar que la descripción es útil y no intrusiva.
- **Test de regresión de sesgos:** Verificación de que el modelo no ha empeorado en el reconocimiento de objetos o personas de diferentes minorías, siguiendo lo definido en el apartado 8 de esta Model Card.

Criterios para el lanzamiento ("Go/No-Go"). El sistema no se actualiza ni se vende si no cumple estos mínimos:

- **Integridad ante peligros:** 0 falsos negativos ante situaciones críticas (coches, huecos, barreras…) en el dataset de prueba.

- **Calidad de descripción:** Las descripciones deben ser comprensibles y precisas, al nivel de un humano (medido con estándares técnicos como CIDEr >90).
- **Utilidad real:** En pruebas con usuarios ciegos, el sistema debe recibir una calificación de "Excelente" en facilidad de uso (puntuación SUS >80).
- **Comportamiento en incertidumbre:** Si la confianza de detección es <60%, el sistema siempre **debe** activar el protocolo de advertencia ("no estoy seguro").

5. Privacidad y protección de datos

Este sistema se clasifica como de **Alto Riesgo** según la **EU AI Act**, debido al procesamiento de datos biométricos y la afectación a derechos fundamentales.

- **Base legal:** Interés legítimo (Art. 6.1.f GDPR), excepción para accesibilidad (Art. 9.2.g GDPR) y sin necesidad de autorización por derechos de autor (Art. 31 LPI).
- **Minimización de datos:** Procesamiento 100% local (*on-edge*). Las imágenes se analizan en la memoria volátil del dispositivo y se destruyen tras generar la descripción.

Medidas de protección

1. **Difuminado automático:** Anonimización de rostros y matrículas en el flujo de vídeo antes de realizar la inferencia.
2. **Modo de privacidad reforzada:** Desactivación automática de la cámara en zonas sensibles (baños, vestuarios, hospitales).
3. **Retención nula:** Al apagar el dispositivo o finalizar la sesión, se ejecuta un borrado automático de cualquier dato temporal en la memoria caché.

6. Propiedad intelectual y modelo de apertura

El proyecto adopta un modelo de negocio híbrido **"Open Core + Servicios"**, liberando una parte del código, mientras los ingresos se obtienen a través de servicios y soporte (personalización para entidades, formación para profesionales…).

Componente	Licencia	Justificación
Código Principal	Propietaria	Protege contra usos no éticos (vigilancia masiva) y asegura trazabilidad exigida por la IA Act.
Módulo de Visión	MIT (Open Source)	Permite auditoría externa de la "visión" del sistema y colaboración comunitaria.
Modelo (Pesos)	CC BY-NC	Permite investigación académica pero impide explotación comercial no autorizada.
Documentación	CC BY	Fomenta la transparencia y educación.

Para gestionar el módulo de visión, se asumen los siguientes compromisos con la comunidad disponibles en el repositorio: Código de Conducta y Guía de Contribución. Además, nos comprometemos a mantener este módulo actualizado y público.

7. Riesgos, límites y gobernanza del sistema

Principales riesgos identificados

1. **Privacidad de terceros:** Grabación accidental en espacios públicos. *Mitigación:* Procesamiento local efímero y difuminado por defecto.
2. **Fallo crítico de seguridad física:** Fallo al detectar un peligro inminente. *Mitigación:* Formación obligatoria de que el sistema es un complemento, no un sustituto del bastón.
3. **Viabilidad técnica en 2029:** Dependencia de la madurez de chips neuromórficos para evitar sobrecalentamiento. *Contingencia:* Plan alternativo de despliegue de modelos reducidos.

Otros aspectos de gran importancia

- **Líneas rojas:** Este sistema estará libre de almacenamiento de imágenes, publicidad, identificación facial, y toma de decisiones para el usuario.
- **Parada del sistema:** El usuario posee la soberanía total. Pulsación larga de 3 segundos para apagar totalmente el sistema.
- **Canal de reclamaciones:** Gestión telefónica en colaboración con la ONCE con respuesta garantizada en 72h. Rendición de cuentas por parte de Cegados por la IA (empresa desarrolladora).

8. Usuarios, sesgos y colectivos afectados

El sistema cuenta con **usuarios directos**, que serán personas con discapacidad visual (grupo vulnerable que gana autonomía pero corre riesgo de dependencia tecnológica o aislamiento si el sistema falla), y **terceros no usuarios**, es decir, ciudadanía general que es grabada pasivamente.

Colectivos vulnerables

- **Por renta:** Un alto coste podría excluir a personas con bajos recursos, creando una "brecha". *Mitigación:* Acuerdos de subvención con la ONCE y administración pública.
- **Por raza/etnia:** Las personas racializadas corren mayor riesgo de ser mal identificadas o no detectadas por sesgos heredados en los datasets de visión por computador (subrepresentación de pieles oscuras).
- **Por entorno (rural/urbano):** El sistema, entrenado mayoritariamente con datos urbanos, podría tener un rendimiento inferior en zonas rurales o con señalética atípica, dejando a estos usuarios con una asistencia de menor calidad.

Interseccionalidad: El riesgo se multiplica en cruces de categorías. En estos casos, la suma de barreras económicas, la menor precisión técnica del modelo por su etnia y la falta de datos de su entorno pueden dejar al usuario en una situación de desprotección crítica.

Como **fuentes** de sesgos se encuentran los datos, por infrarrepresentación de objetos o situaciones culturales, y el producto, por barreras de precio o complejidad de la interfaz. Se han identificado riesgos de **sesgo interpretativo**, donde el sistema podría describir el entorno basándose en estereotipos culturales o raciales, por ello se aplica:

- **Límite de personalización:** Se prohíbe explícitamente la personalización ideológica. El usuario **no** puede configurar el sistema para recibir descripciones filtradas por prejuicios.
- **Neutralidad:** El modelo está alineado para **no emitir juicios de valor**.
    - *Incorrecto:* "Esa persona parece peligrosa".
    - *Correcto:* "Persona caminando rápido hacia ti".
- **Justicia interseccional:** Se presta especial atención en la fase de prueba a usuarios que pertenezcan a colectivos vulnerables para asegurar que el sistema no falla desproporcionadamente en su reconocimiento o trato.

Como **métricas de equidad**, se establecen mínimos para el pase a producción: **paridad de recall**, ya que la diferencia en la tasa de detección de peatones entre diferentes grupos étnicos no debe superar el 5%, y **validación cruzada**, pues el sistema debe superar pruebas de usabilidad en entornos culturalmente diversos antes de cada actualización.

9. Limitaciones, dudas abiertas y pasos futuros

Limitaciones conocidas

- **Hardware:** La duración de la batería en procesamiento intensivo de vídeo sigue siendo el principal cuello de botella físico. Los avances tecnológicos de cara a 2029 serán de gran importancia, tanto para la batería como para el chip.
- **Contextos complejos:** El sistema puede tener dificultades en situaciones de caos visual, donde se recomendará al usuario confiar en los métodos tradicionales, pues no se pretende sustituirlos.

Pasos futuros

- **Postura “macro”:** La organización mantiene una postura intermedia entre el *escepticismo* (vigilancia de riesgos) y el *safetyism* (problemas sociales y económicos), alejándose de un aceleracionismo que pueda ser peligroso.
- **Hoja de ruta:** Se realizará una investigación continua en compresión de modelos para reducir requisitos de hardware, y una ampliación de los acuerdos con organizaciones para subvencionar el dispositivo y evitar la brecha de acceso económico.

Anexos que complementan esta Model Card

- Anexo A - Arquitectura Técnica y Flujo de Datos
- Anexo B - Datos, Entrenamiento y Métricas de Calidad
- Anexo C - Análisis de Privacidad y Cumplimiento Normativo
- Anexo D - Matriz de Riesgos y Filosofía Ética
- Anexo E - Modelo "Open Core" y Estrategia de Propiedad Intelectual