AUDITORÍA PARTICIPATIVA Y ABIERTA DE SESGOS INTERSECCIONALES EN MODELOS DE IA VISUAL

Sofia Llàcer Caro (CVC), Laura Martín Montañez (CVC), Lluís Gómez Bigordà (CVC-UAB),
Ágata Bañón Pérez (Radical Data), Gala Pin Ferrando, Mireia Orra de Salsas, Eva Cruells Lopez, Sonia Ruiz
Margarita Padilla, Marta Cruells López, Nadia Nadesan

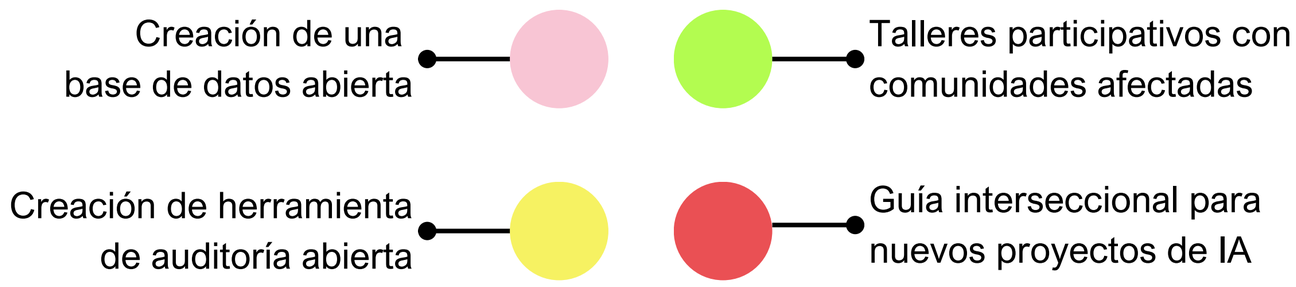
1. PROBLEMÁTICA

Los sistemas de inteligencia artificial (IA) que trabajan con imágenes aprenden de millones de datos y, al hacerlo, absorben y amplifican sus sesgos: invisibilizan ciertos cuerpos, refuerzan estereotipos y simplifican identidades complejas.

Las herramientas existentes para evaluar estos sesgos apenas tienen en cuenta que tenemos identidades múltiples y cruzadas (interseccionalidad). Están diseñadas exclusivamente por equipos técnicos, sin involucrar a quienes sufren esos sesgos. El resultado: métricas socialmente insuficientes, y comunidades afectadas excluidas del proceso de evaluación.

2. INTERVISIONS

InterVisions investiga cómo y por qué los sistemas de IA que relacionan texto e imágenes producen representaciones sesgadas, y construye herramientas abiertas junto a comunidades afectadas para detectarlo. Nuestras actividades principales se centran en elaborar directrices abiertas para la evaluación de estos sesgos en futuros proyectos con IA, consultando las comunidades afectadas como base del conocimiento para esta evaluación.



3. CIENCIA ABIERTA Y ENFOQUE PARTICIPATIVO

6 talleres participativos

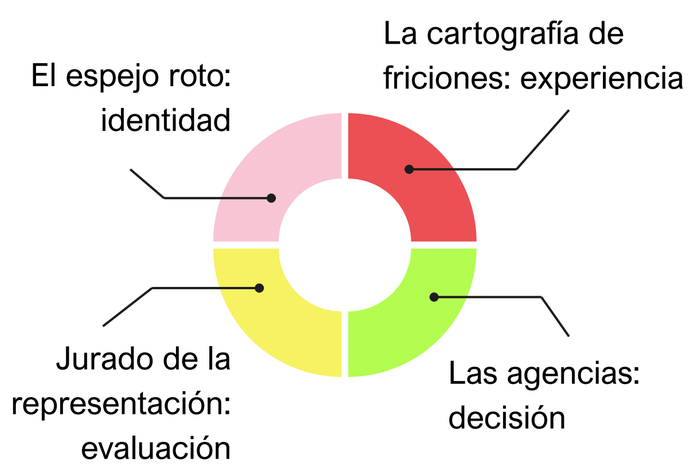
3 países diferentes

20 participantes por taller

Habrá un portal de evaluación público donde cualquier organización pueda comparar la equidad de distintos modelos y publicar sus propios resultados. La base de datos elaborada se publicará con licencia abierta y documentada para ser reutilizada.

Los materiales de los talleres se publicarán también en abierto para que otras organizaciones puedan replicarlos. Las guías de impacto interseccional estarán disponibles para futuros proyectos. Los resultados se diseminarán en GitHub y HuggingFace.

Actividades de los talleres



4. CRONOGRAMA

Otoño 2025: Diseño de la guía interseccional e investigación técnica de evaluación de modelos

Primavera 2026: Ejecución de los talleres participativos, anotación de datos

Invierno 2026: Diseño del formato y organización de los talleres participativos

Verano 2026: Análisis de resultados y retorno a las comunidades participantes

5. TRANSFERENCIA

Los canales de transferencia incluyen talleres replicables con sociedad civil y colectivos afectados, con materiales en abierto; participación en la Festa de la Ciència de Barcelona con actividades para público general sobre sesgos en IA; un portal público para acceder al estándar de evaluación y publicar resultados; y diseminación académica en revistas y congresos de IA y ética (p. ej., RightsCon).

Este trabajo es presentado por el Centre de Visió per Computador (CVC) en el marco de su compromiso con una IA abierta y socialmente responsable.

Queremos agradecer de antemano el entusiasmo y la colaboración de los colectivos con los que estamos llevando a cabo esta labor.

Co-funded by the European Union under the GA no. 101214711. CERV programme (Call: CERV-2024-CHAR-LITI-CHARTER). The content reflects only the authors' views; the European Union is not responsible for any use that may be made of the information it contains.