# INTERACCION

Revista Digital de AIPO

Asociación Interacción Persona-Ordenador

Vol. 6, No 2 (2025)

INTERACCION
Revista Digital de AIPO
Asociación Interacción Persona-Ordenador
Vol. 6, No 2 (2025)
ISSN 2695-6578

# Comité Editorial

**ISSN electrónico:** 2695-6578

**Editado en:** Asociación Interacción Persona-Ordenador (AIPO)
C/ María de Luna, 1, Universidad de Zaragoza, Departamento de Informática e Ingeniería de Sistemas, edificio Ada Byron, 50018 – Zaragoza,
aipo@aipo.es

**Año de edición:** 2025



**Publicado por:** Asociación Interacción Persona-Ordenador (AIPO)
C/ María de Luna, 1, Universidad de Zaragoza, Departamento de Informática e Ingeniería de Sistemas, edificio Ada Byron, 50018 – Zaragoza,
aipo@aipo.es

INTERACCION
Revista Digital de AIPO
Asociación Interacción Persona-Ordenador
Vol. 6, No 2 (2025)
ISSN 2695-6578

INTERACCION
Revista Digital de AIPO
Asociación Interacción Persona-Ordenador
Vol. 6, No 2 (2025)
ISSN 2695-6578

# Preámbulo

Los sistemas interactivos influyen en todos los aspectos de la vida de las personas, asistimos a una continua evolución de los paradigmas clásicos de interacción a nuevas formas de interactuar, es esencial investigar y compartir el conocimiento de estos paradigmas emergentes. Con este espíritu trabaja la Asociación Interacción Persona-Ordenador (AIPO) desde hace 20 años.

La revista Interacción, revista digital de la Asociación Interacción Persona-Ordenador (AIPO), nace con este número 1 con el objetivo de difundir el conocimiento de la Interacción Persona-Ordenador (IPO) y servir de vínculo entre los científicos y profesionales que desarrollen actividades en este ámbito, y con la finalidad de potenciar la transferencia de sus resultados a la sociedad.

La IPO es un campo de investigación multidisciplinario, por ello, la revista presenta contribuciones del ámbito de la Informática como: usabilidad, el diseño centrado en el usuario, accesibilidad, experiencia de usuario, juegos serios, computación ubicua, realidad aumentada, realidad virtual, computación móvil y desarrollo de interfaces de usuario, pero además, se quiere fortalecer la publicación de trabajos de investigación en áreas de diseño industrial, robótica, psicología, etc. relacionadas con la IPO.

Esta revista se distribuye a todos los socios, así mismo, se defiende que su publicación sea de acceso abierto que fomente el avance del conocimiento científico a disposición de todos, por ello su contenido es libremente accesible por Internet.

La revista Interacción selecciona los artículos para publicar en un sistema de revisión por pares, doble ciego, siguiendo las buenas prácticas de las revistas académicas. Es una revista enfocada a la comunidad en España e Iberoamericana y publica artículos en español. Interacción se publica en formato exclusivamente digital, con una periodicidad semestral, publicándose dos números al año. La llamada de artículos está abierta todo el año.

**Resumen del Volumen 6, Número 2:**
Este número consta de siete artículos, que plantean contribuciones interesantes y de actualidad en diversos temas. Los robots sociales, y sus diversas aplicaciones, se abordan en dos artículos diferentes, desde puntos de vista diversos. Así, Cubero et al. plantean una plataforma de evaluación para robots sociales asistenciales orientados a personas mayores, mientas que Mendoza et al. abordan el uso de este tipo de dispositivos desde el punto de vista de la computación afectiva, integrando aspectos de inteligencia artificial y modelos grandes de lenguaje. Precisamente la inteligencia artificial generativa, tan en auge ahora, está presente en muchos artículos de este número. Así, Lara et al. han utilizado inteligencia artificial para trabajar con texto y pictogramas en una plataforma para el apoyo a personas con trastorno del espectro autista. Por su parte, Amengual-Alcover et al. presentan un conjunto de datos de expresiones faciales generado con modelos de difusión y pensado para mejorar la investigación en el reconocimiento de expresiones faciales. También dentro de la inteligencia artificial, Gaya-Morey et al. han trabajado con el concepto de inteligencia artificial explicable, realizando un estudio comparativo entre varios modelos, dentro del dominio de clasificación de actividades en vídeo.

Finalmente, el número se completa con otros dos artículos de otros ámbitos de la IPO. Vázquez-Ingelmo et al. presentan una propuesta, elaborada con principios de diseño centrado en el usuario, para la recogida de variables clínicas que suponga un avance con respecto al uso tradicional de hojas de cálculo. Por su parte, Li & Macías Iglesias realizan un trabajo de revisión y propuesta para la automatización de la evaluación de la usabilidad con la técnica de *Thinking Aloud*, manejando diversas tecnologías.

Nuestro agradecimiento a todas las personas que han contribuido en este número, tanto a nivel de autoría como en las diversas labores de edición y revisión que tan importantes son para que la revista alcance los estándares de calidad esperados.

Jesús Gallardo, Daniel Guasch, Raquel Hervás, Cristina Manresa-Yee y Lourdes Moreno
Editores de Interacción, Revista digital de AIPO

INTERACCION
Revista Digital de AIPO
Asociación Interacción Persona-Ordenador
Vol. 6, No 2 (2025)
ISSN 2695-6578

# Tabla de contenidos

Diseño y accesibilidad en PlanTEA-WM: una plataforma con IA para planificar rutinas para personas con TEA
*José Lara Navarro, Ana Isabel Molina Díaz y Carmen Lacave Rodero* ..... *6*

Enfoques arquitectónicos en plataformas de evaluación *Wizard of Oz* para robots sociales: el caso de SHARA-WoZ
*Guillermo Cubero, Laura Villa y Ramón Hervás* ..... *22*

Más allá de las hojas de cálculo: creando flujos para la definición, validación e interoperabilidad de variables clínicas
*Andrea Vázquez-Ingelmo, Islem Román Nieto-Campo, Alicia García-Holgado, Francisco José García Peñalvo, Antonio Sánchez-Puente y Pedro L. Sánchez* ..... *33*

Interacción natural y emocional con robots sociales
*Liany Mendoza, Eva Cerezo, Loreto Matinero, Adrián Arribas y Sandra Baldassarri* ..... *44*

UIBAIFED: Un dataset de expresiones faciales generado por IA para visibilizar la diversidad
*Esperança Amengual-Alcover, Maria Francesca Roig-Maimó, Ramon Mas-Sansó y Miquel Mascaró-Oliver* ..... *58*

Evaluación efectiva de usabilidad mediante técnicas de análisis y extracción de conocimiento
*Shuoshuo Li y José Antonio Macías Iglesias* ..... *70*

Exploración de las preferencias de usuarios expertos sobre las regiones de importancia como explicaciones en clasificación de actividades en vídeo
*F. Xavier Gaya-Morey, Jose M. Buades-Rubio, Scott MacKenzie, Raquel Lacuesta y Cristina Manresa-Yee* ..... *84*

INTERACCION
Revista Digital de AIPO
Asociación Interacción Persona-Ordenador
Vol. 6, No 2 (2025)
ISSN 2695-6578

# Diseño y accesibilidad en PlanTEA-WM: una plataforma con IA para planificar rutinas para personas con TEA

# Design and accessibility in PlanTEA-WM: an AI-based platform for planning routines for people with ASD

**José Lara Navarro**
Escuela Superior de Informática
Universidad de Castilla-La Mancha
Ciudad Real, España
jose.lara3@alu.uclm.es

**Ana Isabel Molina Díaz**
Escuela Superior de Informática
Universidad de Castilla-La Mancha
Ciudad Real, España
anaisabel.molina@uclm.es

**Carmen Lacave Rodero**
Escuela Superior de Informática
Universidad de Castilla-La Mancha
Ciudad Real, España
carmen.lacave@uclm.es

Recibido: 01.10.2025 | Aceptado: 23.10.2025

## Palabras Clave

Tecnologías asistivas
Trastorno del Espectro Autista
Planificación y anticipación
Pictogramas
Inteligencia Artificial
Diseño Centrado en el Usuario
Usabilidad
Accesibilidad

## Resumen

PlanTEA-WM es una plataforma web colaborativa que permite planificar y anticipar rutinas para personas con Trastorno del Espectro Autista (TEA) mediante pictogramas. Surge como evolución de la aplicación móvil PlanTEA, superando sus limitaciones de portabilidad y la ausencia de capacidades multiusuario. El sistema soporta dos roles principales: planificador y planificado, lo que permite la gestión compartida de rutinas en contextos familiares, educativos y clínicos. La plataforma integra funcionalidades clave, como un calendario de eventos, importación y exportación de planificaciones, un buscador de pictogramas, así como un traductor de texto-a-pictogramas apoyado por modelos de Inteligencia Artificial (IA) generativa. Desde sus primeras fases, el diseño ha estado guiado por principios de accesibilidad (WCAG 2.2) y pautas de diseño específicas para usuarios con TEA, validadas mediante un proceso iterativo con asociaciones de personas con TEA y familiares, y personal experto. Estas decisiones han garantizado la literalidad, la reducción de la carga cognitiva y la flexibilidad en los modos de visualización e interacción soportados. Los resultados obtenidos han dado lugar a un sistema robusto y usable, que facilita la anticipación y la colaboración entre diferentes agentes implicados en su uso (personas con TEA, cuidadores y familiares). Como trabajo futuro, se plantea la incorporación de mecanismos de gestión de imprevistos, una IA más adaptativa y su generalización a otros colectivos de usuarios con necesidades de anticipación, planificación y apoyo visual estructurado.

## Keywords

Assistive technologies
Autism Spectrum Disorder
Planning and anticipation
Pictograms
Artificial Intelligence
User-Centered Design
Usability
Accesibility

## Abstract

PlanTEA-WM is a collaborative web platform designed to support the planning and anticipation of routines for people with Autism Spectrum Disorder (ASD) through pictograms. It evolves from the original mobile app PlanTEA, addressing its limitations of single-device use and lack of multiuser features. The system defines two main roles: planner and planned user, enabling shared management of routines across family, educational, and clinical contexts. Key functionalities include an event calendar, import/export of routines, a pictogram search engine connected to the ARASAAC API, and a text-to-pictogram translator enhanced by generative AI models. From its early stages, the design was guided by accessibility principles (WCAG 2.2) and ASD-specific heuristics, validated through iterative collaboration with expert associations. These guidelines ensured literal representation, reduced cognitive load, and flexible visualization modes. Results show a robust and usable system that strengthens anticipation and multi-role collaboration. Future work focuses on implementing mechanisms to manage unexpected events, developing more adaptive AI, and extending the platform to other groups who can benefit from structured visual routines.

## 1. Introducción

El Trastorno del Espectro Autista (TEA) es un trastorno del neurodesarrollo que se manifiesta desde edades tempranas y que presenta una gran heterogeneidad en sus formas de expresión. Entre sus características más comunes se encuentran las dificultades en la comunicación e interacción social, así como la preferencia por actividades y acciones repetitivas, estructuradas y predecibles (Lord et al., 2018). En este contexto, los apoyos visuales, y en particular los Sistemas Aumentativos y Alternativos de Comunicación (SAAC), constituyen un recurso esencial para facilitar la comunicación, la planificación y anticipación de actividades, lo cual reduce la incertidumbre y fomenta la autonomía de las personas con TEA (Chia et al., 2018).

Los pictogramas, integrados dentro de los SAAC, se han consolidado como un medio eficaz para representar rutinas y transmitir instrucciones de forma clara, sencilla y accesible (Morales-Hidalgo et al., 2018). Su empleo resulta especialmente útil en entornos clínicos, familiares y educativos. Por su parte, la planificación y anticipación de actividades permiten reducir la incertidumbre y la ansiedad ante cambios inesperados y promueve la comprensión y la participación activa de las personas con TEA en la organización de sus tareas cotidianas (Neimy et al., 2022).

PlanTEA (Hernández et al., 2022), una aplicación móvil desarrollada inicialmente para dispositivos Android, surge para facilitar la planificación y anticipación de actividades cotidianas para personas con TEA. Contempla dos roles: el *planificador*, encargado de crear y gestionar las rutinas, y la persona con TEA, o *planificado*, que puede anticipar y seguir paso a paso las tareas representadas en secuencias de pictogramas (planificaciones). Incluye un editor de planificaciones empleando técnicas de *drag-and-drop*[1], y un cuaderno de comunicación[2] aumentativa, concebido como un tablero interactivo de pictogramas que permite la expresión y comunicación en distintos contextos.

Aunque dicha aplicación obtuvo una valoración muy positiva en sus primeras evaluaciones y pruebas de uso, realizadas en colaboración con varias asociaciones de personas con TEA (Hernández et al., 2022; Valencia et al., 2024), presenta una serie de limitaciones: solo se puede usar en un único dispositivo, la información se almacena en local, lo que impide su sincronización, y no admite ser usado por varios usuarios a la vez. Estas restricciones dificultan su adopción en contextos en los que la colaboración entre los diferentes agentes implicados (familiares, terapeutas y profesionales) resulta fundamental.

Con el objetivo de superar estas carencias, se ha desarrollado PlanTEA-WebMultiuser (PlanTEA-WM), una plataforma *web* que evoluciona la aplicación original hacia un sistema multiusuario, multirrol y accesible desde cualquier dispositivo y navegador. Desde sus primeras fases, el proyecto ha contado con la colaboración activa de asociaciones especializadas en TEA, de ámbito local, regional y nacional (AUTRADE[3], FACLM[4] y FESPAU[5]), lo que ha permitido orientar y adaptar las decisiones de diseño a las necesidades reales de los usuarios finales en entornos clínicos, educativos y familiares. Este trabajo conjunto ha sido clave para validar prototipos, priorizar funcionalidades y garantizar la usabilidad y accesibilidad de la plataforma desarrollada, tal y como se detallará en secciones posteriores.

Por tanto, el objetivo de este artículo es presentar PlanTEA-WM desde una perspectiva centrada en el diseño y la accesibilidad, profundizando en los aspectos que hacen de la plataforma una herramienta diferenciadora: la aplicación de principios de Diseño Centrado en el Usuario (DCU), la incorporación de heurísticas de usabilidad específicas para usuarios con TEA, la validación iterativa con personas expertas en el dominio perteneciente a distintas asociaciones, así como la integración de técnicas de Inteligencia Artificial (IA) generativa[6] para asistir en el proceso de creación de las planificaciones visuales. Asimismo, se exponen los avances logrados y las líneas de evolución futura, mostrando cómo la tecnología puede adaptarse a las particularidades de este colectivo y contribuir a mejorar su calidad de vida (Valencia et al., 2019).

[1] La técnica de *drag-and-drop* es un mecanismo de interacción en entornos de interfaces gráficas de usuario (GUI) que permite la manipulación directa de objetos mediante la acción de seleccionar (*drag*) un elemento con un dispositivo apuntador, desplazarlo sobre la superficie de la interfaz, y liberarlo (*drop*) en una ubicación destino, desencadenando una operación asociada.

[2] Un *cuaderno de comunicación* es un recurso de apoyo basado en símbolos o pictogramas que facilita la expresión y comprensión del lenguaje en personas con dificultades de comunicación, tal y como ocurre con personas con TEA.

[3] AUTRADE (Asociación Regional de Personas con Autismo y Otros Trastornos del Neurodesarrollo): https://autrade.info/

[4] FACLM (Federación Autismo Castilla-La Mancha): https://www.autismocastillalamancha.org/

[5] FESPAU (Federación Española de Autismo): https://fespau.es/

[6] La *Inteligencia Artificial Generativa* (IAG) hace referencia a un conjunto de técnicas de aprendizaje automático capaces de modelar patrones complejos en grandes volúmenes de datos y generar contenido nuevo (texto, imágenes, audio, entre otros) que mantiene coherencia con dichos patrones.

## 2. Descripción general de PlanTEA-WM

PlanTEA-WM es una plataforma *web* diseñada para apoyar la planificación y anticipación de rutinas mediante el uso de pictogramas. Su diseño se basa en un enfoque multiusuario y multirrol, lo que permite que distintos agentes implicados, como familiares, terapeutas o profesionales (de contextos educativos o clínicos), colaboren en la creación y gestión de rutinas visuales para personas con TEA.

Al igual que en PlanTEA, PlanTEA-WM contempla dos roles principales. Por un lado, el ***planificador*** es el encargado de crear y gestionar las *planificaciones* y los *eventos* asociados. Una *planificación* es una secuencia de pictogramas que representa una rutina o actividad, mientras que un *evento* corresponde a la asignación de una planificación concreta a una fecha y hora determinadas en el calendario. Por otro lado, está el rol de *planificado*, que corresponde a la **persona con TEA,** quien accede a una vista, en formato de reproductor, que le permite consultar y seguir las planificaciones paso a paso de manera visual y estructurada. Esta separación de roles garantiza que cada perfil disponga de una experiencia adaptada a sus necesidades específicas. La Figura 1 muestra el *panel de control* disponible para el usuario *planificador*, en el que se muestra la lista de usuarios *planificados* por dicho usuario, con los eventos asociados a cada uno de ellos.

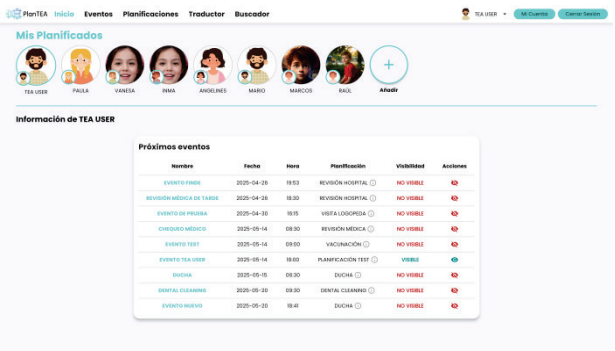



*Figura 1. Panel de control del usuario planificador en PlanTEA-WM, en el que se muestra la lista de personas con TEA planificadas y los próximos eventos de cada una. La interfaz permite consultar información relevante (fecha, hora y planificación vinculada al evento mediante una vista previa), así como seleccionar qué planificación será visible para el usuario planificado.*

PlanTEA-WM resuelve las limitaciones presentes en la aplicación original: integra un **sistema multiusuario y multirrol**, que permite a un mismo usuario *planificador* gestionar a varios usuarios *planificados*, y, a su vez, compartir la responsabilidad de un mismo usuario con TEA entre varios *planificadores*, sin duplicar datos; incorpora un **traductor de texto-a-pictogramas (de la base de datos de ARASAAC[7])**, basado en herramientas de IA, que facilita, en parte, el proceso de diseño y creación de las planificaciones; y admite la **importación y exportación de planificaciones en formatos estándar** (PDF y XML), lo que favorece la portabilidad de la información y su uso en diferentes contextos.

En cuanto a su arquitectura, PlanTEA-WM se compone de un *frontend* desarrollado en Flutter Web[8], siguiendo el patrón Modelo-Vista-VistaModelo (MVVM), lo que facilita la separación entre la interfaz, la lógica de presentación y la gestión de estado. El *backend*, implementado en Node.js/Express[9], bajo el patrón Modelo-Vista-Controlador (MVC), se encarga de la lógica de negocio y de la comunicación con la base de datos. Para la persistencia de la información se emplea PostgreSQL[10], que asegura escalabilidad e integridad en la gestión de *planificadores*, *planificados*, *planificaciones* y *eventos* (Fielding et al, 2000). También soporta la integración con servicios externos, como la API de ARASAAC, los Grandes Modelos de Lenguaje, a través de OpenRouter[11], y la plataforma Brevo[12], para el envío de invitaciones por correo electrónico. La seguridad se refuerza mediante autenticación con *tokens* y control de permisos en función de los roles definidos.

Cabe destacar que algunas funcionalidades presentes en la *app* original, como el cuaderno de comunicación, fueron descartadas en PlanTEA-WM. Esta decisión se tomó tras un proceso de diseño participativo y colaborativo (Villamin et al., 2024) con las asociaciones expertas implicadas y siguiendo una metodología iterativa e incremental basada en la construcción progresiva de prototipos (Bjarnason et al., 2023), que permitió priorizar aquellas funcionalidades consideradas más útiles y esenciales para soportar la anticipación y gestión de rutinas.

Gracias a esta combinación de funcionalidades, PlanTEA-WM ofrece una solución accesible y colaborativa, que puede ser utilizada desde cualquier navegador y dispositivo conectado a Internet. Una demostración práctica del uso de PlanTEA-

[7] ARASAAC (Centro Aragonés para la Comunicación Aumentativa y Alternativa), repositorio abierto de pictogramas y otros recursos visuales ampliamente utilizado en sistemas SAAC. Enlace: https://arasaac.org/
[8] Flutter Web: https://flutter.dev/multi-platform/web
[9] Node.js/Express: https://expressjs.com/
[10] PostgreSQL: https://www.postgresql.org/
[11] OpenRouter es un agregador de APIs que permite acceder múltiples LLMs a través de una única interfaz. Enlace: https://openrouter.ai/
[12] Brevo es una plataforma *software* que ofrece herramientas de *marketing* y ventas, como envío de *emails*, automatización o campañas de SMS. Enlace: https://www.brevo.com/es/

WM se encuentra disponible en https://youtu.be/DcT5Q7XFnOQ.

## 3. Diseño Centrado en el Usuario

El desarrollo de PlanTEA-WM ha seguido, desde sus primeras etapas, un enfoque DCU, en el que han participado asociaciones de personas con TEA (AUTRADE, FACLM y FESPAU). Estas entidades han aportado su experiencia práctica en contextos clínicos, educativos y familiares.

La metodología, **iterativa e incremental**, se ha basado en la construcción progresiva de prototipos. Se han diseñado prototipos de baja y alta fidelidad, en Balsamiq[13] (Figura 2) y Figma[14] (Figura 3), respectivamente. Además, se creó un prototipo funcional inicial con un *mini-backend* en Flask[15], que sirvió como "prueba de concepto" antes de la migración definitiva. Cada iteración incluyó sesiones de validación con terapeutas y profesionales de las asociaciones colaboradoras, lo que permitió definir y ajustar, de forma temprana, los requisitos y las funcionalidades, reduciendo así posibles errores.

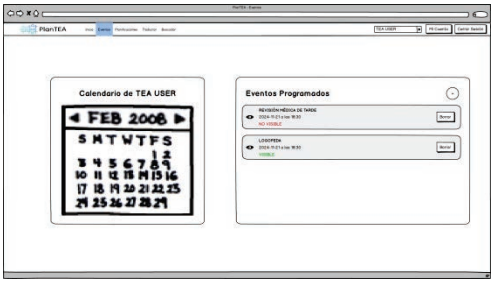



*Figura 2. Mockup de baja fidelidad en Balsamiq de la vista de calendario de eventos en PlanTEA-WM.*

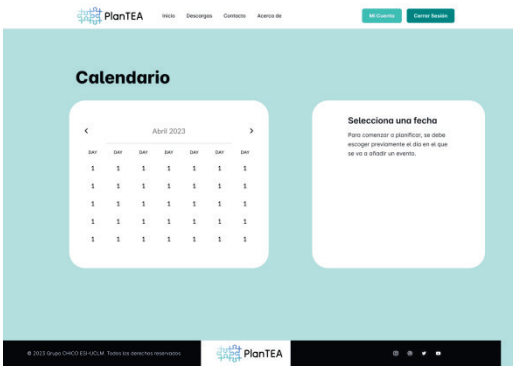



*Figura 3. Diseño de alta fidelidad en Figma de la vista de calendario de eventos en PlanTEA-WM.*

[13] Balsamiq: https://balsamiq.com/
[14] Figma: https://www.figma.com
[15] Flask: https://flask.palletsprojects.com/en/stable/

Las reuniones y jornadas de trabajo con las personas expertas de las asociaciones han permitido tomar decisiones clave. En la *I Jornada de Tecnologías para la inclusión de personas con TEA*[16] (Ciudad Real, 15 de enero de 2024), se decidió acotar el contexto de uso de la aplicación al de las consultas médicas, dejando en principio de lado otros contextos o escenarios de uso. Además, se decidió eliminar el cuaderno de comunicación, puesto que se consideró menos prioritario, al existir soluciones más completas y dinámicas en el mercado (Logan et al., 2017). Meses más tarde, en la reunión con el personal técnico de la asociación nacional FESPAU, se confirmaron los requisitos iniciales, se planteó la posibilidad de integrar técnicas de IA generativa para dar soporte a la creación de planificaciones y se discutieron algunas propuestas de mejora relacionadas con la usabilidad y accesibilidad de la aplicación. Posteriormente, durante la presentación de PlanTEA-WM[17] al presidente de la Confederación de Autismo España[18], en el laboratorio de usabilidad del grupo CHICO[19] de la Escuela Superior de Informática de la Universidad de Castilla-La Mancha, se valoró muy positivamente la incorporación de la IA generativa y su potencial para ahorrar tiempo en la generación de rutinas. Por último, en la *II Jornada de Trabajo: Tecnología para la inclusión de personas con TEA*[20] se mostraron versiones avanzadas del sistema desarrollado, incorporando ya la IA generativa, y se discutieron nuevas propuestas de ampliación funcional.

La Tabla 1 resume los principales cambios de diseño derivados del *feedback* recibido en estas sesiones.

*Tabla 1. Cambios de diseño y funcionalidades derivados del feedback de las asociaciones colaboradoras.*

| Necesidad detectada | Cambio aplicado | Resultado |
|---|---|---|
| Evitar sobrecarga funcional en la fase inicial | Eliminación del cuaderno de comunicación | Foco en planificaciones y rutinas, con una interfaz más simple y usable |
| Contextualizar el sistema en un dominio concreto | Validación de prototipos en escenarios de consultas médicas | Ajuste de la plataforma a un contexto realista y relevante |

[16] https://esi.uclm.es/index.php/2024/01/19/jornadas-de-tecnologias-para-la-inclusion/
[17] https://blog.uclm.es/grupochico/presentacion-de-plantea-al-presidente-de-la-confederacion-autismo-espana/
[18] Autismo España: https://autismo.org.es/
[19] CHICO (Computer Human Interaction and Collaboration): https://blog.uclm.es/grupochico/
[20] https://www.autismocastillalamancha.org/participacion-en-la-ii-jornada-de-trabajo-de-la-catedra-uclm-telefonica-tecnologia-para-la-inclusion/

| Necesidad detectada | Cambio aplicado | Resultado |
|---|---|---|
| Facilitar la creación rápida de rutinas visuales | Desarrollo del traductor de texto-a-pictogramas | Conversión automática de frases en secuencias de pictogramas, reduciendo tiempos de creación |
| Procesar materiales textuales extensos | Funcionalidad de traducción de documentos a pictogramas | Simplificación del trabajo con textos largos o ya preparados |
| Incluir apoyos más avanzados en la creación de rutinas | Integración de la IA generativa | Facilidad y rapidez en la creación de planificaciones y mayor apoyo a *planificadores* |
| Centralizar la información de cada usuario con TEA en la vista del *planificador*. | Creación de un *dashboard*[21] con próximos eventos y planificaciones asociadas | Información clave en un único espacio, facilitando la supervisión y el control |
| Facilitar la comprobación rápida de rutinas asociadas | Inclusión de vista previa de planificaciones en el panel de control | Posibilidad de revisar de un vistazo la secuencia de pictogramas sin acceder a la vista completa |
| Dificultad para localizar y gestionar varias planificaciones de un mismo usuario con TEA | Incorporación del patrón *Master-Detail*[22] en el panel de control | Mayor claridad y rapidez en la gestión, reduciendo pasos intermedios |
| Necesidad de personalización con referentes familiares | Opción de sustituir pictogramas por fotos reales en determinadas rutinas | Mayor identificación por parte de la persona con TEA, facilitando la familiaridad de los contenidos y la anticipación |
| Ajustar la representación de pictogramas a diferentes perfiles de usuarios con TEA | Inclusión de tres modos de visualización configurables: texto + imagen, solo texto o solo imagen | Mayor flexibilidad y personalización de la interfaz, facilitando la adaptación a distintos niveles de comprensión |

Además de los cambios aplicados, el proceso de co-diseño ha permitido identificar **patrones de interfaz que resultaban inadecuados** para el público objetivo y que, por tanto, fueron descartados: interfaces densas, con múltiples paneles visualizados simultáneamente en pantalla, que incrementaban la carga cognitiva de la persona con TEA al requerir demasiada atención dividida; menús desplegables con niveles de anidamiento excesivos, ya que generaban confusión durante la navegación y dificultaban la localización de opciones (Zhang et al., 2025); metáforas visuales abstractas, que podían resultar poco claras y difícilmente reconocibles; descartar animaciones llamativas o transiciones visuales complejas, ya que podían generar distracción o saturación (Uitdenbogerd et al., 2022). Frente a estas alternativas, se priorizó un diseño lineal, literal y predecible, en el que cada pictograma o elemento visual mantiene un único significado explícito y consistente. Por ejemplo, se incorporó el icono de una papelera para eliminar planificaciones creadas y para suprimir pictogramas de una planificación; y el del lápiz para modificar la imagen por defecto de un pictograma por otra cualquiera que eligiese el usuario. Esta elección resulta coherente con las heurísticas específicas para diseños orientados a personas con TEA, y que se discuten en la siguiente sección.

Como parte del proceso de validación de PlanTEA-WM, se llevó a cabo una evaluación preliminar con expertos y profesionales del ámbito del TEA (Lara et al., 2025). Esta evaluación se realizó durante un seminario *online* organizado por la asociación AUTRADE, en marzo de 2025, en el que participaron 44 personas, principalmente profesionales que trabajan directamente con personas con TEA (terapeutas y educadores) y familiares.

Los resultados de esta evaluación fueron muy positivos. En cuanto a la *utilidad percibida de las funcionalidades* soportadas por PlanTEA-WM, todas obtuvieron puntuaciones medias superiores a 4,5 sobre 5, destacando especialmente la colaboración multiusuario (4,75), la traducción de texto a pictogramas (4,68) y las sugerencias generadas en el modo asistido por la IA (4,68). Respecto a la valoración de la aceptación y adopción tecnológica, medida mediante las dimensiones del *framework* TAM (*Technology Acceptance Model*) (Davis, 1989), se obtuvo que la *utilidad percibida* alcanzó una media de 4,48, y tanto la *intención de uso* personal como la intención de recomendar la plataforma obtuvieron medias de 4,55. Los participantes valoraron especialmente la *facilidad de uso*, la integración de IA como elemento diferenciador clave, y la posibilidad de colaboración entre los diferentes actores.

Esta evaluación se centró en valorar la aplicación desde el punto de vista del usuario con rol *planificador.* Dado que PlanTEA-WM diferencia dos roles, con características y

[21] *Dashboard*: Panel de control que centraliza y muestra información clave en una sola vista, facilitando la supervisión y la toma de decisiones.
[22] *Master-Detail*: Patrón de diseño de interfaz que presenta en paralelo un listado de elementos (*master*) y el detalle del elemento seleccionado (*detail*), facilitando la navegación y la gestión de información.

necesidades distintas, y el DCU propone validar el diseño con los potenciales usuarios finales, actualmente se está preparando una evaluación más exhaustiva, en el que participarán ambos. Los resultados de dicha evaluación serán fundamentales para refinar la plataforma y valorar la efectividad y futura adopción de PlanTEA-WM.

## 4. Accesibilidad y heurísticas aplicadas en PlanTEA-WM

La accesibilidad en entornos digitales dirigidos a personas con TEA requiere de un enfoque dual. En el diseño de PlanTEA-WM, por un lado, se han adoptado las **Pautas de Accesibilidad para el Contenido Web (WCAG 2.2)** (World Wide Web Consortium, 2023) como referencia internacional para garantizar la perceptibilidad, operabilidad y comprensibilidad de la plataforma. Por otro, se han aplicado **heurísticas de diseño específicas para aplicaciones dirigidas a usuarios con TEA** (Valencia et al., 2021; Valencia et al., 2022; Aguiar et al., 2020), desarrolladas a partir de literatura previa y validadas de manera iterativa con asociaciones colaboradoras, con el fin de responder a las necesidades cognitivas concretas de los usuarios con TEA.

### 4.1. Heurísticas específicas para personas con TEA

El desarrollo de la plataforma se ha apoyado en principios derivados de la literatura sobre neurodiversidad (Shuck et al., 2024) y del proceso iterativo de validación con las asociaciones colaboradoras. Entre ellos destacan:

- **Literalidad:** cada pictograma o icono tiene un único significado explícito, evitando metáforas abstractas o confusas.
- **Consistencia y previsibilidad:** la posición y el comportamiento de los elementos permanecen estables en toda la interfaz.
- **Reducción de carga cognitiva:** simplificación de la estructura visual, navegación lineal y supresión de menús anidados en exceso.
- **Flexibilidad de representación:** inclusión de tres modos de visualización de pictogramas (texto+imagen, solo texto, solo imagen), adaptables a diferentes perfiles de usuario.
- **Apoyos familiares:** posibilidad de sustituir pictogramas por fotografías reales, aumentando la identificación y el reconocimiento del contenido (Hartley et al., 2015).
- **Prevención de distracciones:** eliminación de animaciones llamativas y transiciones complejas, priorizando la claridad y la estabilidad.

Estos principios han guiado todo el proceso de desarrollo de PlanTEA-WM, desde el prototipado hasta la implementación, actuando como criterios de decisión en el diseño y ayudando a descartar patrones que podían incrementar la carga cognitiva de la aplicación.

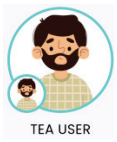
### 4.2. Aplicación de las WCAG 2.2 en PlanTEA-WM

La aplicación de las WCAG 2.2 en PlanTEA-WM se ha materializado tanto en aspectos perceptuales, como de contraste y color, así como de operabilidad, consistencia y claridad en la interacción. A continuación, se detallan los principales aspectos implementados y aquellos que permanecen pendientes de desarrollo o revisión.

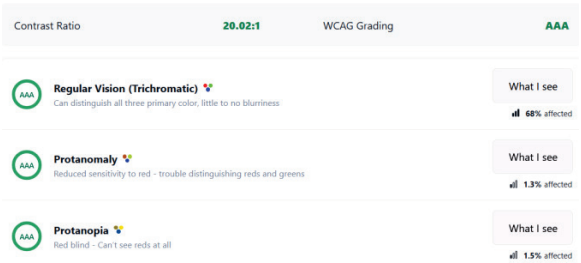
#### *4.2.1. Criterios implementados*

Se ha aplicado el criterio **1.4.3 (Contraste mínimo, nivel AA)**, asegurando que todos los textos y botones presenten una relación de contraste superior a 4.5:1. En varios casos, como en los nombres de usuarios *planificados* en el *dashboard* del *planificador*, o en los títulos de las tarjetas de eventos, se alcanzaron ratios muy superiores (16:1), lo que garantiza un alto grado de legibilidad. Asimismo, se empleó el criterio **1.4.6 (Contraste mejorado, nivel AAA)** en elementos clave, como botones de acción y títulos, favoreciendo la máxima visibilidad en diferentes dispositivos.

La Figura 4 muestra un ejemplo de validación de contraste aplicado a los nombres de los usuarios *planificados*, donde se alcanzó un ratio de 20:02:1, muy por encima de los niveles exigidos por la normativa.




*(a) Visualización original del nombre con su avatar.*




*(b) Resultados del análisis con la herramienta WhoCanUse*

*Figura 4. Validación de contraste aplicado al nombre de un usuario planificado en el dashboard de PlanTEA-WM.*

La aplicación del criterio **1.4.1 (Uso del color)** evita que la codificación de la información dependa exclusivamente del color. Así, por ejemplo, los estados de los botones se refuerzan mediante etiquetas textuales o iconos. Del mismo

modo, la plataforma es completamente operable por teclado, cumpliendo con los criterios **2.1.1 (Teclado)** y **2.1.2 (Sin trampas de teclado)**, lo que permite la gestión de usuarios, eventos y planificaciones sin necesidad de usar el ratón.

En relación con la comprensión de la interfaz, se cumplió con el criterio **2.4.6 (Encabezados y etiquetas)**, empleando descripciones claras y consistentes en campos de formularios y secciones principales. Además, los formularios cumplen el criterio **3.3.1 (Identificación de errores)**, puesto que usan mensajes breves y lenguaje sencillo, evitando tecnicismos que pudieran dificultar su interpretación. Finalmente, el criterio **3.1.5 (Nivel de lectura)** también se ha tenido en cuenta, con el empleo de frases cortas y vocabulario cotidiano, lo cual favorece tanto a los *planificadores* como a las personas con TEA.

### *4.2.2. Criterios pendientes*

A pesar de los avances y mejoras realizadas, algunos criterios de las WCAG 2.2 aún requieren atención. El criterio **1.1.1 (Contenido no textual)** se cumple parcialmente, ya que no se han incorporado descripciones alternativas de manera sistemática para todos los pictogramas y fotografías personalizadas. Algo similar ocurre con el criterio **2.4.4 (Propósito de los enlaces)**, dado que algunos iconos, como el de papelera o el de edición, carecen todavía de un texto alternativo explícito que facilite su interpretación por parte de tecnologías de asistencia.

El criterio **2.2.1 (Ajuste de tiempo)** tampoco está implementado en su totalidad, ya que la aplicación no permite configurar ni desactivar la duración de la visualización de las rutinas durante su reproducción. Finalmente, es necesario reforzar el cumplimiento del criterio **4.1.2 (Nombre, rol y valor)**, mediante una revisión exhaustiva de atributos ARIA[23], con el fin de garantizar que todos los elementos interactivos sean reconocibles de forma adecuada por lectores de pantalla.

## 4.3. Validación cromática y pruebas de percepción

El contraste en el contenido visualizado en la aplicación se ha evaluado mediante ratios y pruebas de percepción, bajo diferentes condiciones de visión cromática. Se han empleado simuladores para verificar cómo se perciben los elementos en casos de **protanopia** (ausencia de rojos), **deuteranopia** (ausencia de verdes) o **tritanopia** (ausencia de azules), así como en sus variantes de percepción reducida. La Figura 5 muestra una representación de la interfaz en estas condiciones, confirmando que la paleta de colores seleccionada mantiene la diferenciación suficiente entre texto y fondo en escenarios de deficiencia cromática. Estas validaciones aseguran que la experiencia no dependa de una visión estándar, reforzando la robustez del diseño en contextos reales.

*(a) Vista normal*

*(b) Protanopia*

*(c) Deuteranopia*

*(d) Tritanopia*

*Figura 5. Validación cromática de la interfaz de planificaciones de PlanTEA-WM en condiciones de visión alterada: (a) visión estándar, (b) protanopía, (c) deuteranopia y (d) tritanopia.*

[23] ARIA (*Accessible Rich Internet Applications*): Conjunto de atributos definidos por el W3C que mejoran la accesibilidad de las aplicaciones *web*, permitiendo que tecnologías de Asistencia como los lectores de pantalla interpreten correctamente el rol, nombre y estado de los elementos de la interfaz.

En resumen, la integración de las WCAG 2.2, junto con la aplicación de heurísticas específicas de diseño de *software* dirigido a usuarios con TEA, ha permitido establecer un marco de accesibilidad consistente en PlanTEA-WM, garantizando así que la plataforma responda a necesidades reales y reduzca posibles barreras de uso. Con ello, la accesibilidad se consolida como un criterio transversal del sistema y se establece una base sólida para futuras evaluaciones en entornos clínicos, familiares y educativos.

## 5. Inteligencia Artificial

La incorporación de técnicas de IA generativa en PlanTEA-WM responde a la necesidad de reducir el tiempo y la carga de trabajo de los *planificadores* en la creación de rutinas visuales. A diferencia de la aplicación original, que exigía añadir manualmente cada pictograma, la nueva plataforma integra un traductor de texto-a-pictogramas, asistido por modelos de lenguaje de gran escala (LLM)[24]. Esta funcionalidad permite generar de manera rápida una propuesta de planificación que debe revisarse por la persona planificadora antes de ser presentadas a la persona con TEA (Tam et al., 2024).

### 5.1. Flujo de integración técnica

El flujo de integración de la funcionalidad de traductor de texto-a-pictogramas basado en IA generativa en PlanTEA-WM se inicia con la entrada textual (consulta o *prompt*[25]) proporcionada por el *planificador*. Este texto se envía al *backend* junto con un *pre-prompt* restrictivo (Chen et al., 2025), diseñado para guiar al modelo hacia la generación de respuestas compuestas por instrucciones breves, literales y sin metáforas. El *backend* gestiona la comunicación con el modelo de lenguaje a través de OpenRouter[26], empleando fragmentos o *streams* de comunicación que permiten recibir la respuesta de manera incremental. Gracias a este enfoque, el planificador visualiza en el *frontend* cómo la propuesta se va generando paso a paso, reduciendo así la percepción de latencia y facilitando la interacción en tiempo real.

Una vez recibida la secuencia completa, el *frontend* aplica procesos de *limpieza de partículas* y *lematización* (Jurafsky, n.d.), antes de solicitar los pictogramas asociados a la API de ARASAAC**:**

- **Limpieza de partículas:** se eliminan elementos lingüísticos que no aportan significado visual ni son representables en pictogramas. Por ejemplo, conectores como "entonces", "luego" o artículos como "el", "la" no se traducen en pictogramas y podrían saturar innecesariamente la secuencia. Así, la frase "*Luego el niño se lava las manos*" se simplifica a "*niño lavar manos*" para la búsqueda de pictogramas, ya que el texto que los acompaña se mantiene sin cambios.
- **Lematización:** se transforma cada palabra a su forma canónica o "lema", lo que mejora la correspondencia con los pictogramas disponibles. Por ejemplo, el verbo *"lavándose"* se convierte en *"lavar"*, o el sustantivo en plural *"zapatos"* pasa a *"zapato"*. De este modo, la búsqueda en ARASAAC es más eficaz, al aumentar la probabilidad de encontrar un pictograma coincidente.

Tras esta etapa, el sistema activa el módulo de búsqueda de pictogramas, a través de la localización en la base de datos de ARASACC del pictograma correspondiente a cada término procesado. Finalmente, los resultados se muestran en el editor de planificaciones, en el que el *planificador* podrá revisar, sustituir o eliminar pictogramas antes de almacenar la planificación definitiva.

### 5.2. Evaluación preliminar de LLMs

De forma preliminar, y a la espera de un estudio más amplio con profesionales y personas con TEA, se ha realizado una comparación preliminar de diferentes LLMs, con el fin de seleccionar una opción adecuada para su integración operativa en PlanTEA-WM. Esta evaluación ha tenido un carácter exploratorio y su objetivo ha sido analizar la viabilidad del uso de la IA en la generación de planificaciones visuales, para reducir así los tiempos de creación y garantizar la coherencia de las secuencias propuestas.

Los modelos comparados fueron **DeepSeek V3[27],** que constituye un modelo de gran escala basado en una arquitectura de tipo *Mixture of Experts* (MoE), que permite seleccionar dinámicamente expertos especializados en diferentes contextos, logrando así una generación de texto de alta calidad y con tiempos de inferencia ajustados

[24] Los *modelos de lenguaje de gran escala* (*Large Language Models*, LLM) son sistemas de aprendizaje profundo entrenados con grandes volúmenes de datos textuales para predecir y generar secuencias lingüísticas coherentes. Constituyen una de las principales aplicaciones de la Inteligencia Artificial Generativa (IAG), al posibilitar la producción automática de texto con fines comunicativos, creativos o analíticos.

[25] En el ámbito de la Inteligencia Artificial Generativa (IAG), un *prompt* es la instrucción o conjunto de entradas textuales que guían a un modelo para producir una salida específica, condicionando el contenido, el estilo o el formato de la generación.

[26] OpenRouter: https://openrouter.ai/

[27] Deepseek V3: https://api-docs.deepseek.com/news/news1226

(Rajbhandari et al., 2022); **Gemma 3 1B IT**[28]**,** un modelo ligero, diseñado para ser eficiente en coste y latencia, adecuado para escenarios en los que se prioriza la obtención de una respuesta rápida frente a la complejidad del procesamiento (Google DeepMind, 2025).; y **LLaMA 3.3 70B Instruct**[29]**,** por ser un referente de gran tamaño en el estado del arte, ser ampliamente empleado en tareas de generación instructiva y por su capacidad para producir respuestas detalladas y consistentes (Wei et al., 2022). La selección de estos tres modelos ha buscado cubrir diferentes rangos de complejidad, escalabilidad y eficiencia, asegurando una visión comparativa amplia en esta fase inicial.

La evaluación de las tres opciones seleccionadas se ha llevado a cabo mediante la simulación de **cinco situaciones representativas** del uso de la plataforma PlanTEA-WM, seleccionadas por su relevancia en contextos clínicos, educativos y familiares. Estas situaciones, recogidas en la Tabla 2, incluyen: (i) una rutina de autonomía personal, (ii) una acción funcional en la sociedad, (iii) una planificación con recompensa, (iv) una situación de autorregulación emocional y (v) una interacción social o petición de ayuda. Cada uno de estos *prompts* se procesó con los tres modelos a testear, bajo condiciones homogéneas de generación, con el fin de asegurar comparabilidad entre las salidas.

Para valorar los resultados se ha diseñado una **rúbrica** ***ad-hoc*****,** que incluye cinco criterios de comparación: (i) claridad del lenguaje, (ii) adecuación de la respuesta a personas con TEA, es decir, el uso de un estilo sencillo, literal y accesible, (iii) estructuración de la respuesta en secuencias paso a paso, (iv) relevancia de la respuesta respecto al contexto del escenario planteado y (v) adecuación al tipo de *prompt* y a su finalidad comunicativa. Cada criterio se ha evaluado en una escala de Likert de 1 a 5, con descripciones precisas de los distintos niveles de desempeño. Además de la puntuación global, derivada de la aplicación de la rúbrica, se han registrado métricas de rendimiento, incluyendo el tiempo inicial de respuesta y el tiempo total necesario hasta disponer de una planificación utilizable tras aplicar correcciones menores.

Los resultados de esta evaluación exploratoria ponen de manifiesto una tensión entre calidad y eficiencia. Tal y como se muestra en la Figura 6, los modelos de mayor escala producen salidas con una estructura más clara y adaptada a las necesidades de las personas con TEA, mientras que el modelo más ligero destaca por su rapidez en los tiempos de respuesta, aunque con una menor consistencia en la calidad de las planificaciones generadas. En este contexto, **DeepSeek V3** ofrece un equilibrio especialmente favorable, combinando altos niveles de adecuación en la rúbrica con tiempos competitivos, lo que ha motivado su elección como modelo provisional a integrar en PlanTEA-WM. No obstante, es necesario insistir en que el análisis comparativo realizado es preliminar y está limitado a un conjunto reducido de escenarios representativos. Es necesario, por tanto, realizar un estudio más exhaustivo y formal, actualmente en preparación, que incorpora la participación de profesionales y personas con TEA, para validar estos hallazgos en contextos reales.

*Tabla 2. Situaciones representativas utilizadas en la evaluación preliminar de modelos de IA para su integración en PlanTEA-WM, con sus descripciones en español e inglés.*

| Tipo de situación | Español | English |
|---|---|---|
| Rutina de autonomía personal | Cómo cepillarme los dientes | *How to brush my teeth* |
| Acción funcional en la sociedad | Qué hacer para coger un autobús | *What to do to catch a bus* |
| Planificación con recompensa | Planificación para ir al hospital y después al parque | *Planning to go to the hospital and then to the park* |
| Autorregulación emocional | Cómo calmarme cuando estoy enfadado | *How to calm down when I'm angry* |
| Interacción social / Petición de ayuda | Cómo pedir ayuda en la escuela | *How to ask for help at school* |

## 6. Escenarios de uso ilustrativos

Esta sección presenta tres escenarios de uso que ejemplifican la utilidad de PlanTEA-WM en contextos clínicos, familiares y educativos. En todos ellos se emplean las vistas principales de la plataforma (panel de control del *planificador*, gestor de planificaciones, calendario de eventos y reproductor en la vista del *planificado*), así como los módulos de apoyo (buscador de pictogramas, traductor de texto y documentos a pictogramas en el modo asistido por IA). Además, se pone de manifiesto las características multiusuario y multirrol soportadas por el sistema, mostrando cómo una misma persona *planificada* puede ser gestionada de forma coordinada por distintas personas *planificadoras* sin duplicación de datos.

[28] Gemma 3: https://ai.google.dev/gemma/docs/core?hl=es-419
[29] LLaMA 3.3 70B Instruct: https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct

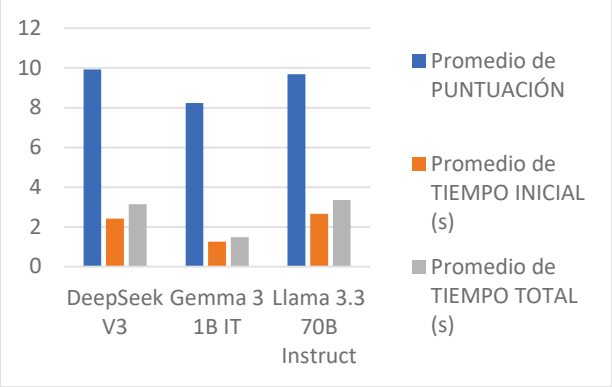



*Figura 6. Evaluación comparativa preliminar de modelos de IAG (DeepSeek V3, Gemma 3 1B IT y LLaMA 3.3 70B Instruct) a integrar en PlanTEA-WM. Se representan los valores medios de la rúbrica de valoración diseñada, expresada sobre 10, junto con el tiempo inicial de respuesta y el tiempo total hasta la obtención de una planificación utilizable.*

### 6.1. Preparación coordinada de una consulta médica

Se parte de la siguiente situación ficticia: "*María, terapeuta de un niño con TEA, llamado Pablo, y Javier, padre del menor, necesitan preparar y anticipar la visita al pediatra y asegurar su correcta realización el día de la cita.*"

El proceso comienza accediendo al ***dashboard* del *planificador***, en el que María selecciona la tarjeta de usuario *planificado* (Pablo) y la opción "Compartir planificado". A continuación, introduce la dirección de correo electrónico de Javier. El sistema valida que dicha dirección pertenece a un usuario registrado en PlanTEA-WM, generando una invitación que se envía de forma automática a su dirección de correo electrónico (véase Figura 7a, que muestra la tarjeta de Pablo en el *dashboard* con la acción de compartir, y Figura 7b, que representa la ventana emergente en la que se introduce la dirección de correo).

Javier recibe un mensaje en su correo electrónico con un enlace a PlanTEA-WM. Al acceder, aparece un cuadro de diálogo que le permitirá aceptar o rechazar la invitación. Una vez aceptada, Javier adquiere permisos de co-gestión del perfil de Pablo, sin que se produzca duplicación de datos ni necesidad de crear un nuevo usuario. La Figura 7c muestra este diálogo de confirmación en la plataforma.

Con la colaboración establecida, Javier podrá abrir el **editor de planificaciones** para crear la rutina denominada "Visita al pediatra". En esta vista podrá arrastrar pictogramas, reorganizar pasos y, cuando lo estime conveniente, sustituir algunos de ellos por fotografías reales y familiares para el niño, que refuercen la identificación de dichos elementos por parte de Pablo (Figura 8).

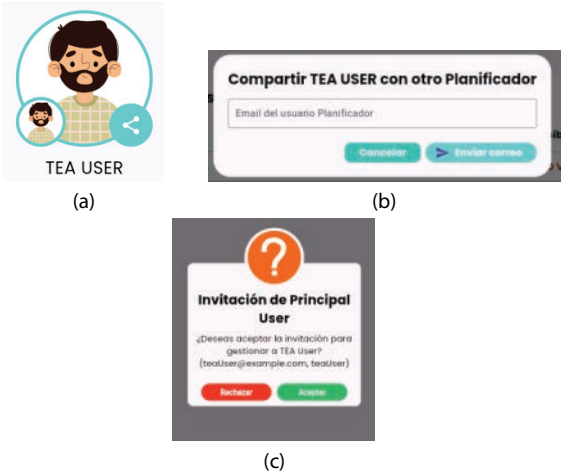



*Figura 7. Proceso de compartición de un usuario planificado en PlanTEA-WM. (a) Selección de la tarjeta del usuario con TEA desde el dashboard del planificador. (b) Ventana emergente para introducir la dirección de correo electrónico del planificador invitado. (c) Diálogo de confirmación mostrado en la plataforma al aceptar la invitación, mediante el cual el nuevo planificador adquiere permisos de co-gestión sin duplicar datos.*


*Figura 8. Ventana modal de edición de pictogramas que permite modificar el nombre y la imagen asociada.*

Una vez creada la planificación, Javier accede a la ventana de eventos y programa la "Visita al pediatra" para la fecha y hora de la cita, asociándola a la planificación anteriormente creada y marcándola como visible para su hijo. La Figura 9 muestra el aspecto del calendario, con el formulario de creación del evento, que incluye los campos de título, fecha/hora, planificación vinculada y conmutador de visibilidad.

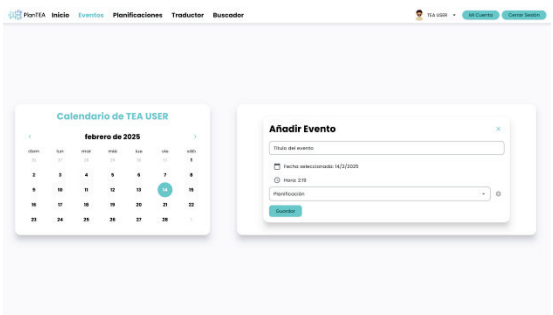



*Figura 9. Calendario de PlanTEA-WM con el formulario para añadir eventos, que permite definir título, fecha y hora*

Cuando María revisa en el *dashboard los* próximos eventos de Pablo, observa la planificación asociada al nuevo evento y decide introducir algunos pasos que Javier ha omitido y que considera importantes para el niño. Para ello, activa el **traductor asistido por IA** e introduce el siguiente *prompt*: "*preparar visita al pediatra con espera y revisión*". El *backend* procesa la entrada y devuelve una propuesta inicial de secuencia paso a paso, que posteriormente se podrá ajustar manualmente. La Figura 10 muestra la interfaz del traductor en modo asistido por la IA, en el que se puede ver el área de entrada de la consulta o *prompt* (parte superior), así como la respuesta generada y su representación en formato de secuencia de pictogramas (parte central de la interfaz).

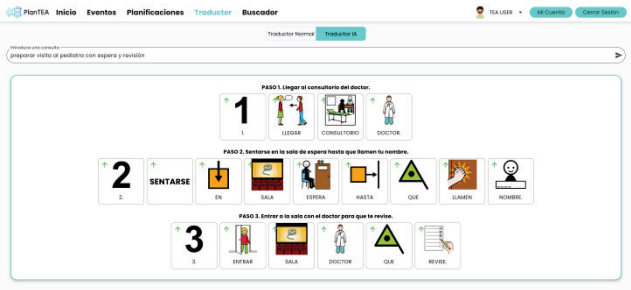



*Figura 10. Vista del traductor en modo asistido por la IA, en el que, a partir de un texto de entrada o prompt, se genera una rutina acompañada de su representación en pictogramas.*

El día señalado, Pablo accederá al reproductor de la planificación, disponible para el rol *planificado*, pudiendo avanzar paso a paso por la rutina diseñada de manera clara y sencilla.

La aplicación también permite poder seguir y simular dicha planificación con antelación al día del evento, permitiendo así anticipar y preparar en casa dicha situación.

El escenario de uso descrito ilustra, de forma explícita, la utilidad del carácter multiusuario y colaborativo soportado por PlanTEA-WM, así como la integración del traductor con IA, el editor de planificaciones y la programación de eventos en el calendario, todos ellos elementos que contribuyen a una anticipación estructurada y coordinada.

### 6.2. Traducción de las normas del aula y planificación semanal en un entorno educativo

En un contexto escolar, el tutor y la logopeda comparten la responsabilidad de asistir a una alumna con TEA. Para favorecer su comprensión de las dinámicas del aula, el tutor utiliza el **traductor de documentos** de PlanTEA-WM para convertir el texto del documento "Normas del aula" a su representación como secuencia de pictogramas.

La aplicación procesará todas las palabras del documento y buscará su correspondencia visual en la base de datos de pictogramas de ARASAAC. El resultado de este proceso se muestra en la Figura 11, en la que se puede ver la salida generada por el traductor de documentos con la secuencia visual resultante.

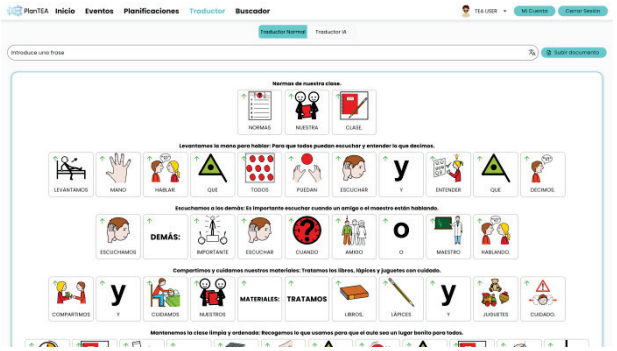



*Figura 11. Resultado del traductor de documentos en PlanTEA-WM, que convierte un texto de "Normas del aula" en una secuencia completa de pictogramas obtenidos de ARASAAC.*

Posteriormente, la logopeda quiere elaborar varias planificaciones temáticas, como "Entrada al aula", "Recreo" o "Salida". De este modo, cada alumna o alumno con TEA accede a estas planificaciones en el **modo de visualización** configurado para su perfil, ya sea combinando texto e imagen, utilizando solo pictogramas o únicamente texto. Esta flexibilidad garantiza que cada persona pueda seguir las rutinas según sus necesidades específicas de comunicación y comprensión.

Con las planificaciones definidas, el tutor programa **eventos diarios** en el calendario para cada franja horaria de la semana escolar, activando como **visible** únicamente la rutina que corresponde a cada momento del día. De este modo, se evita la sobrecarga de estímulos y se refuerza la anticipación de manera situada en el tiempo. Finalmente, las planificaciones creadas pueden **exportarse a PDF**, para ser colocadas en el aula en formato físico, o a **XML**, para generar copias de seguridad o permitir la interoperabilidad con otros sistemas.

Este escenario de uso muestra cómo el traductor de documentos agiliza la traducción y reutilización de materiales ya existentes y cómo la combinación de planificaciones configurables, eventos y visibilidad proporciona un apoyo situacional eficaz para la anticipación en entornos escolares.

### 6.3. Autonomía en los desplazamientos

Para ilustrar este escenario, se plantea la siguiente situación: "*Javier, padre de Pablo, un niño con TEA, desea preparar la*

*rutina "Ir en autobús para acudir a terapia" con el fin de fomentar su autonomía en los desplazamientos",* actividad cotidiana que puede resultar muy desafiante y estresante para ambos (Afif et al., 2022). Para agilizar la tarea de planificación y anticipación de esta rutina, usando PlanTEA-WM, se utilizará el **traductor asistido por IA**, introduciendo el *prompt*: *"ir en autobús desde casa a terapia, comprar billete, esperar, bajar en parada correcta"*. El sistema procesa la entrada y devuelve una propuesta inicial de pasos, traducidos a pictogramas de ARASAAC, que Javier podrá editar y reorganizar manualmente utilizando el **editor de planificaciones**.

Con el objetivo de mejorar la identificación de los elementos que componen la secuencia, Javier sustituye algunos pictogramas críticos, como "esperar en la parada" o "cruzar la calle", por fotografías reales del entorno físico en el que se realizará la tarea (parada de autobús, calle, …). Pablo visualizará esta planificación en el **modo de visualización**, disponible en su perfil, lo que asegura la adecuación a su nivel de comprensión. En la Figura 12, se pueden ver los diferentes modos de visualización que Pablo podría tener configurados.

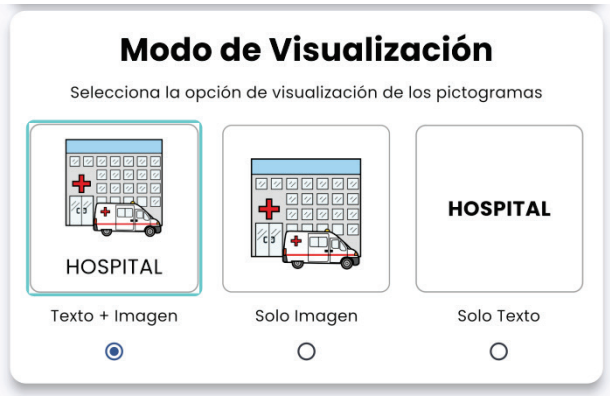



*Figura 12. Modos de visualización de pictogramas disponibles en PlanTEA-WM.*

Antes del día de la actividad, su hermana Laura edita el evento "Ir en autobús a terapia" en el calendario. En este caso, introduce un cambio para reflejar un **imprevisto**: la necesidad de bajarse una parada antes debido a obras en la vía. En el formulario de edición del evento se sustituye el pictograma representativo de la parada final por otro alternativo, manteniendo el evento marcado como visible para Pablo. Cuando llega la hora del desplazamiento, Pablo accede al **reproductor de la planificación**, disponible en el rol *planificado*, en el que consulta la rutina "Ir en autobús", que es la que le han marcado como visible.

Este escenario de uso pone de relieve la importancia de contemplar mecanismos para la **gestión de imprevistos** en PlanTEA-WM. Actualmente, la modificación de eventos permite realizar sustituciones manuales, pero resulta necesario desarrollar un sistema que marque explícitamente la diferencia entre la planificación original y la modificada, mostrando simultáneamente el pictograma previsto inicialmente y el nuevo. Este tipo de representación ayudaría a las personas con TEA a comprender que ha habido un cambio respecto a lo esperado, minimizando la incertidumbre y facilitando la aceptación de la variación (Rydzewska, 2016). La inclusión de esta capacidad se plantea como una línea de trabajo futuro esencial para reforzar la robustez de la herramienta en contextos cotidianos en los que los imprevistos son frecuentes.

## 7. Perspectivas de evolución

La mejora de PlanTEA-WM contempla una hoja de ruta que combina la consolidación de las funcionalidades ya implementadas con la incorporación de nuevas capacidades, que amplíen su alcance y robustez. El objetivo es avanzar hacia una plataforma más adaptativa, personalizada y transferible a otros colectivos, manteniendo siempre un enfoque centrado en la persona y en la accesibilidad universal.

Entre los retos inmediatos se encuentra el de reducir la dependencia de proveedores externos en la ejecución de los modelos de IA generativa, dado que ello introduce riesgos de latencia y disponibilidad. Una línea de investigación prioritaria será explorar despliegues en entornos *on-premise* o *edge*[30], que aporten mayor control sobre la infraestructura y aumenten la fiabilidad del sistema (Singh et al., 2023).

En paralelo, se identifican varias líneas de mejora diferenciadas:

- **Cobertura en ARASAAC:** persisten casos en los que determinados conceptos no disponen de pictograma asociado. Para mitigar esta limitación, se plantea explorar fuentes visuales alternativas y permitir la incorporación de pictogramas personalizados, garantizando que la representación gráfica cubra un mayor espectro de situaciones.
- **Control automático de calidad:** se prevé incorporar mecanismos de verificación que detecten metáforas, ambigüedades o pasos excesivamente complejos antes de mostrar la

[30] El término *on-premise* hace referencia al despliegue de los modelos en infraestructuras locales bajo control institucional, mientras que *edge* alude a la ejecución en dispositivos cercanos al usuario, con el fin de reducir la latencia y la dependencia de servicios en la nube.

propuesta al *planificador*, reforzando así la usabilidad de las secuencias generadas.

- **Especialización por dominios:** el uso de plantillas de *prompt* específicas (por ejemplo, higiene, salud o educación) permitirá guiar la generación de rutinas de forma más coherente y contextualizada, aumentando la pertinencia de los resultados en cada ámbito de aplicación.

Como siguiente hito, se prevé la realización de estudios controlados con profesionales y personas con TEA, lo que permitirá validar de manera más robusta la utilidad de la funcionalidad basada en IA generativa en contextos reales. Estos ensayos aportarán información empírica sobre su aplicabilidad y servirán de base para la mejora de los mecanismos de supervisión y edición por parte de los planificadores.

A corto plazo, se prioriza contemplar los criterios de accesibilidad aún pendientes, así como la integración de un mecanismo específico para la gestión de imprevistos en las rutinas.

En una segunda fase, se reforzará la integración de técnicas de IA más robustas y adaptativas, capaces de detectar errores y generar propuestas ajustadas a las necesidades particulares de cada usuario. Para ello, se prevé recoger información, tanto explícita (por ejemplo, valoraciones directas sobre la utilidad de una planificación) como implícita (tiempo de permanencia en la aplicación, repeticiones de simulaciones o interacciones más frecuentes) que permitan crear un "perfil" de usuario. Estos datos permitirán un aprendizaje progresivo que facilite la generación de recomendaciones personalizadas y mejor adaptadas a cada usuario particular (Gheewala, 2025). Además, se desarrollará un *chatbot* integrado en la plataforma, para asistir al planificador en la creación de rutinas, automatizar tareas frecuentes y facilitar la interacción mediante lenguaje natural (Cao et al., 2023).

En el medio plazo, el sistema avanzará hacia una personalización más profunda, incluyendo la adaptación automática de la longitud y complejidad de las rutinas, así como la integración con tableros de AAC y sistemas de entrada/salida por voz. Estas funcionalidades reforzarán la utilidad de la plataforma en contextos clínicos, educativos y familiares, aportando flexibilidad y ampliando los canales de interacción disponibles para las personas con TEA.

El proceso de diseño continuará guiado por un enfoque DCU, en el que el *feedback* continuo de asociaciones y profesionales especializados permitirá priorizar y validar las nuevas funcionalidades antes de su despliegue.

Finalmente, en el largo plazo, se contempla la extensión de PlanTEA-WM a otros colectivos más allá del TEA, como personas con afasia, personas adultas con deterioro cognitivo o con discapacidad intelectual. La modularidad de la arquitectura y la flexibilidad de la plataforma hacen posible esta ampliación, que permitirá generalizar y transferir los beneficios de la anticipación y el empleo de rutinas visuales a un espectro más amplio de potenciales usuarios, lo que aumentará el impacto esperado de PlanTEA-WM en la sociedad.

La Figura 13 resume gráficamente este plan de evolución, mediante un *roadmap* en seis hitos, que reflejan la progresión planificada, desde la consolidación inmediata hasta la expansión futura hacia nuevos colectivos.


*Figura 17. Roadmap de evolución de PlanTEA-WM. Principales hitos para la consolidación, extensión y generalización de la plataforma.*

## 8. Conclusiones

PlanTEA-WM constituye la evolución de una aplicación móvil previa (PlanTEA) hacia una plataforma *web* colaborativa, accesible y multirrol, que permite la gestión integral de

rutinas visuales basadas en pictogramas en contextos familiares, educativos y clínicos. La incorporación de un editor de planificaciones, un calendario de eventos y la integración con la API de ARASAAC, para la traducción de texto y documentos, ha permitido superar limitaciones previas, favoreciendo la coordinación entre los diferentes agentes implicados, sin duplicación de datos y garantizando la disponibilidad en cualquier dispositivo conectado a Internet.

La **accesibilidad** se ha consolidado como un eje transversal del desarrollo. La aplicación de las **WCAG 2.2** asegura niveles adecuados de contraste, operabilidad por teclado y claridad en la rotulación, mientras que la introducción de **heurísticas de diseño específicas para usuarios con TEA** refuerza la literalidad, la consistencia y la reducción de la carga cognitiva. La flexibilidad en los modos de visualización de la información (texto+imagen, solo texto o solo imagen), junto con la posibilidad de incorporar contenidos (fotografías) familiares, ha demostrado ser esencial para la adaptación y personalización de los contenidos a distintos perfiles. Asimismo, las pruebas de percepción cromática confirman la robustez de la interfaz frente a diferentes condiciones de visión alterada.

El **proceso de co-diseño iterativo** con asociaciones y profesionales especializados ha permitido ajustar prioridades y descartar patrones contraproducentes, consolidando un diseño centrado en la previsibilidad y la simplicidad. De este modo, se han priorizado funcionalidades orientadas a la planificación y la anticipación de actividades, mientras que se han descartado elementos que podían incrementar la carga cognitiva o generar confusión.

En cuanto al empleo de técnicas de **inteligencia artificial**, la integración de modelos de lenguaje de gran escala en el traductor de texto-a-pictogramas presenta un gran potencial para agilizar la creación de rutinas, reduciendo el tiempo de preparación y proporcionando estructuras claras y secuenciales. No obstante, se mantiene la necesidad de supervisión y edición por parte del *planificador*, persisten limitaciones derivadas de la dependencia de proveedores externos, la cobertura incompleta de determinados conceptos en ARASAAC y la ausencia de pruebas y validaciones con expertos y usuarios finales.

Como **líneas de desarrollo futuro**, se prioriza el cierre de los criterios de accesibilidad, abordando los aún pendientes, el diseño de un **mecanismo específico para la gestión de imprevistos**, que permita representar simultáneamente el pictograma original y el modificado, y el avance hacia una **IA más adaptativa**, capaz de ajustar automáticamente las propuestas en función del perfil del usuario. Asimismo, se prevé la integración con tableros AAC, la incorporación de entrada/salida por voz y la exploración de despliegues en infraestructuras locales para reducir la dependencia de servicios externos.

En síntesis, PlanTEA-WM establece una **base sólida y extensible** para la planificación de rutinas visuales, integrando estándares de accesibilidad, principios de diseño centrado en el usuario e innovaciones en inteligencia artificial. Con la validación en entornos reales y la ampliación funcional prevista, la plataforma se proyecta no solo como un apoyo eficaz para personas con TEA, sino también como una solución transferible a otros colectivos con necesidades de anticipación y estructuración de actividades.

## Agradecimientos

Este trabajo ha sido desarrollado en el contexto de los proyectos APTEA (Ref. TED2021-131956B-I00), financiado por MICIU/AEI/10.13039/501100011033 y por la Unión Europea NextGenerationEU/PRTR, de los proyectos PlanTEAAF y TEAcompaño (Refs. 2022-GRIN-34175 and Ref. 2025-GRIN-38489, respectivamente), financiados por la Universidad de Castilla-La Mancha y el Fondo Europeo de Desarrollo Regional (FEDER); y de los convenios de colaboración y transferencia con las asociaciones de atención a personas con TEA: AUTRADE (Ref. 220413CONV), FACLM (Ref. 230405CONV) y FesPAU (Ref. 240437CONV).

## Referencias

Afif, I. Y., Manik, A. R., Munthe, K., Maula, M. I., Ammarullah, M. I., Jamari, J., & Winarni, T. I. (2022). Physiological effect of deep pressure in reducing anxiety of children with ASD during traveling: A public transportation setting. *Bioengineering, 9*(4), 157. https://doi.org/10.3390/bioengineering9040157

Aguiar, Y. P., Galy, E., Godde, A., Trémaud, M., & Tardif, C. (2020). AutismGuide: A usability guidelines to design software solutions for users with autism spectrum disorder. *Behaviour & Information Technology, 41*(11), 1132–1150. https://doi.org/10.1080/0144929X.2020.1856927

Bjarnason, E., Lang, F., & Mjöberg, A. (2023). An empirically based model of software prototyping: A mapping study and a multi-case study. *Empirical Software Engineering, 28*(115), 1–40. https://doi.org/10.1007/s10664-023-10331-w

Cao, C. C., Ding, Z., Lin, J., & Hopfgartner, F. (2023). AI chatbots as multi-role pedagogical agents: Transforming engagement in CS education. *arXiv preprint arXiv:2308.03992*. https://arxiv.org/abs/2308.03992

Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2025). Unleashing the potential of prompt engineering for large language models. *Patterns (New York, N.Y.), 6*(6), 101260. https://doi.org/10.1016/j.patter.2025.101260

Chia, G. L. C., Anderson, A., & McLean, L. A. (2018). Use of technology to support self-management in individuals with autism: Systematic review. *Review Journal of Autism and Developmental Disorders, 5*(2), 142–155. https://doi.org/10.1007/s40489-018-0129-5

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly*, *13*(3), 319-340. https://doi.org/10.2307/249008

Fielding, R. T. (2000). *Architectural styles and the design of network-based software architectures* (Doctoral dissertation, University of California, Irvine). University of California, Irvine. https://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm

Gheewala, S., Xu, S. & Yeom, S. In-depth survey: deep learning in recommender systems—exploring prediction and ranking models, datasets, feature analysis, and emerging trends. *Neural Comput & Applic* **37**, 10875–10947 (2025). https://doi.org/10.1007/s00521-024-10866-z

Google DeepMind. (2025). *Gemma 3 technical report*. arXiv. https://doi.org/10.48550/arXiv.2503.19786

Hartley, C., & Allen, M. L. (2015). Symbolic understanding of pictures in low-functioning children with autism: the effects of iconicity and naming. *Journal of autism and developmental disorders*, *45*(1), 15–30. https://doi.org/10.1007/s10803-013-2007-4

Hernández, P., Molina, A. I., Lacave, C., Rusu, C., & Toledano-González, A. (2022). PlanTEA: Supporting planning and anticipation for children with ASD attending medical appointments. *Applied Sciences, 12*(10), 5237. https://doi.org/10.3390/app12105237

Hervás, R., Francisco, V., Méndez, G., & Bautista, S. (2019). A user-centred methodology for the development of computer-based assistive technologies for individuals with autism. In D. Lamas, F. Loizides, L. Nacke, H. Petrie, M. Winckler, & P. Zaphiris (Eds.), *Human-Computer Interaction – INTERACT 2019* (Lecture Notes in Computer Science, vol. 11746, pp. 75–84). Springer. https://doi.org/10.1007/978-3-030-29381-9_6

Jurafsky, D., & Martin, J. H. (n.d.). *Speech and language processing* (3rd ed., draft). Stanford University. Retrieved September 25, 2025, from https://web.stanford.edu/~jurafsky/slp3/

Lara, J., Lacave, C., & Molina, A. I. (2025). PlanTEA-WM: A Multi-User Web Platform for Routine Planning and Anticipating Everyday Situations in Individuals With Autism Spectrum Disorder. *IEEE Access*, 13, 180523-180538. https://doi.org/10.1109/access.2025.3617152

Logan, K., Iacono, T., & Trembath, D. (2017). A systematic review of research into aided AAC to increase social-communication functions in children with autism spectrum disorder. *Augmentative and Alternative Communication, 33*(1), 51–64. https://doi.org/10.1080/07434618.2016.1267795

Lord, C., Elsabbagh, M., Baird, G., & Veenstra-Vanderweele, J. (2018). Autism spectrum disorder. *The Lancet, 392*(10146), 508–520. https://doi.org/10.1016/S0140-6736(18)31129-2

Morales-Hidalgo, P., Roigé-Castellví, J., Hernández-Martínez, C., Voltas, N., & Canals, J. (2018). Prevalence and characteristics of autism spectrum disorder among Spanish school-age children. *Journal of Autism and Developmental Disorders, 48*(10), 3176–3190. https://doi.org/10.1007/s10803-018-3581-2

Neimy, H., & Fossett, B. (2022). Augmentative and Alternative Communication (AAC) systems. In *Handbook of special education research* (pp. 375–401). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-96478-8_20

Rajbhandari, S., Li, C., Yao, Z., Zhang, M., Aminabadi, R. Y., Awan, A. A., Rasley, J., & He, Y. (2022). DeepSpeed-MoE: Advancing Mixture-of-Experts inference and training to power next-generation AI scale. *arXiv*. https://doi.org/10.48550/arXiv.2201.05596

Rydzewska, E. (2016). Unexpected changes of itinerary – adaptive functioning difficulties in daily transitions for adults with autism spectrum disorder. *European Journal of Special Needs Education*, *31*(3), 330–343. https://doi.org/10.1080/08856257.2016.1187889

Schuck, R. K., & Fung, L. K. (2024). A dual design thinking–universal design approach to catalyze neurodiversity advocacy through collaboration among high-schoolers. *Frontiers in Psychiatry, 14,* 1250895. https://doi.org/10.3389/fpsyt.2023.1250895

Singh, R., & Gill, S. S. (2023). Edge AI: A survey. *Internet of Things and Cyber-Physical Systems*, *3*, 71-92. https://doi.org/10.1016/j.iotcps.2023.02.004

Tam, T. Y. C., et al. (2024). A framework for human evaluation of large language models. *NPJ Digital Medicine, 7*(1), 58. https://doi.org/10.1038/s41746-024-01258-7

Uitdenbogerd, A. L., Spichkova, M., & Alzahrani, M. (2022). Web-based search: How do animated user interface elements affect autistic and non-autistic users? *arXiv*. https://arxiv.org/abs/2211.11993

Valencia, K., Rusu, C., Quiñones, D., & Jamet, E. (2019). The Impact of Technology on People with Autism Spectrum Disorder: A Systematic Literature Review. Sensors (Basel, Switzerland), 19(20), 4485. https://doi.org/10.3390/s19204485

Valencia, K., Rusu, C., & Botella, F. (2021). User Experience Factors for People with Autism Spectrum Disorder. *Applied Sciences*, *11*(21), 10469. https://doi.org/10.3390/app112110469

Valencia, K., Botella, F., & Rusu, C. (2022). A property checklist to evaluate the user experience for people with autism spectrum disorder. In G. Z. Yang (Ed.), *Social computing and social media: Design, user experience and impact. SCSM 2022. Lecture Notes in Computer Science* (Vol. 13335, pp. 205–216). Springer. https://doi.org/10.1007/978-3-031-05061-9_15

Valencia, K., Hernández del Mazo, P., Molina, A. I., Lacave, C., Rusu, C., & Botella, F. (2024). Evaluating PlanTEA: The practice of a UX evaluation methodology for people with ASD. *Universal Access in the Information Society*. Advance online publication. https://doi.org/10.1007/s10209-024-01175-2

Villamin, G. R., & Luppicini, R. (2024). Co-Designing Digital Assistive Technologies for Autism Spectrum Disorder (ASD) Using Qualitative Approaches. *International Journal of Disability, Development and Education*, 1–19. https://doi.org/10.1080/1034912X.2024.2427606

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2022). Finetuned language models are zero-shot learners. *arXiv*. https://doi.org/10.48550/arXiv.2109.01652

World Wide Web Consortium (W3C). (2023). *Web Content Accessibility Guidelines (WCAG) 2.2*. W3C Recommendation. https://www.w3.org/TR/WCAG22/

Zhang, B., Qi, Y., Yang, Y., & Zhang, J. (2025). Research on the interface design of ASD children intervention app based on Kano-entropy weight method. *Frontiers in Psychiatry, 16,* 1508006. https://doi.org/10.3389/fpsyt.2025.1508006

INTERACCION
Revista Digital de AIPO
Asociación Interacción Persona-Ordenador
Vol. 6, No 2 (2025)
ISSN 2695-6578

# Enfoques arquitectónicos en plataformas de evaluación *Wizard of Oz* para robots sociales: el caso de SHARA-WoZ

# Architectural approaches in Wizard of Oz evaluation platforms for social robots: the case of SHARA-WoZ

**Guillermo Cubero**
Departamento de Tecnologías y Sistemas de Información
Universidad de Castilla-La Macha
Ciudad Real, España
Guillermo.Cubero@uclm.es

**Laura Villa**
Departamento de Tecnologías y Sistemas de Información
Universidad de Castilla-La Macha
Ciudad Real, España
Laura.Villa@uclm.es

**Ramón Hervás**
Departamento de Tecnologías y Sistemas de Información
Universidad de Castilla-La Macha
Ciudad Real, España
Ramon.HLucas@uclm.es

Recibido: 06.11.2025 | Aceptado: 03.12.2025

## Palabras Clave

Robots sociales asistenciales
Mago de Oz
Interacción humano-robot
Cuidado geriátrico
Inteligencia artificial conversacional

## Resumen

Los robots sociales asistenciales emergen como solución para el cuidado de personas mayores, pero su validación efectiva requiere metodologías que faciliten la participación de múltiples implicados técnicos y no técnicos. La técnica Wizard of Oz (WoZ) permite evaluar interacciones antes de su implementación autónoma completa, aunque las plataformas actuales carecen frecuentemente de infraestructuras estandarizadas y modulares. Este artículo presenta SHARA-WoZ v2.0.0, una evolución de una plataforma de evaluación para robots sociales asistenciales que aborda sistemáticamente las limitaciones identificadas en su versión anterior. Las contribuciones principales incluyen: (1) migración a una aplicación de escritorio nativa en PyQt6 que reduce la latencia de renderizado en un 42% y el uso de memoria en un 49%; (2) integración de GPT-4o-mini que elimina servicios de traducción intermedios y reduce la latencia conversacional total en un 55%; (3) sistema de generación de respuestas múltiples basadas en estados emocionales que reduce el tiempo de respuesta en modo semi-automático a 2-3 segundos; (4) rediseño modular de la interfaz que facilita la supervisión simultánea de múltiples canales de interacción; y (5) arquitectura orientada a servicios que garantiza escalabilidad y extensibilidad. Los resultados demuestran mejoras cuantificables en rendimiento, latencia y usabilidad, estableciendo una plataforma que equilibra sofisticación técnica con accesibilidad para implicados no técnicos en el desarrollo y evaluación de robots sociales para contextos geriátricos. Estos resultados pueden ser generalizables como modelo arquitectónico efectivo para la aplicación de técnicas de Mago de Oz para el contexto de los Robots Asistenciales y Sociales.

## Keywords

Socially assistive robots
Wizard of Oz
Human-robot interaction
Geriatric care
Conversational artificial intelligence

## Abstract

Socially assistive robots emerge as a solution for elderly care, but their effective validation requires methodologies that facilitate the participation of multiple technical and non-technical stakeholders. The Wizard of Oz (WoZ) method enables the evaluation of interactions before full autonomous implementation, although current platforms frequently lack standardized and modular infrastructures. This article presents SHARA-WoZ v2.0.0, a evolution of an evaluation platform for socially assistive robots that systematically address limitations identified in its previous version. The main contributions include: (1) migration to a native desktop application in PyQt6 that reduces rendering latency by 42% and memory usage by 49%; (2) integration of GPT-4o-mini that eliminates intermediate translation services and reduces total conversational latency by 55%; (3) multiple response generation system based on emotional states that reduces response time in semi-automatic mode to 2-3 seconds; (4) modular interface redesign that facilitates simultaneous supervision of multiple interaction channels; and (5) service-oriented architecture that guarantees scalability and extensibility. Results demonstrate improvements in performance, latency, and usability, establishing a

platform that balances technical sophistication with accessibility for non-technical stakeholders in the development and evaluation of social robots for geriatric contexts. These results can be generalized as an effective architectural model for the application of Wizard of Oz techniques in the context of Assistive and Social Robots.

## 1. Introducción

El envejecimiento progresivo de la población mundial ha impulsado el desarrollo de tecnologías asistenciales orientadas al cuidado de personas mayores. Los robots sociales asistenciales (SARs) emergen como solución prometedora para abordar necesidades de acompañamiento, monitorización de salud y asistencia en actividades diarias (Abdi et al., 2018; Kachouie et al.,2014). Sin embargo, la validación efectiva de estos sistemas requiere de la participación de múltiples implicados (stakeholders), incluyendo usuarios finales, profesionales clínicos, familiares y cuidadores, más que depender exclusivamente de expertos técnicos, lo que incrementa considerablemente la complejidad de coordinación (Tobis et al., 2023). La necesidad de medir resultados sociales, emocionales y terapéuticos demanda marcos de evaluación capaces de acomodar perspectivas de múltiples grupos interesados manteniendo validez científica.

La técnica Wizard of Oz (WoZ) constituye una herramienta fundamental en el desarrollo de sistemas de interacción en general, y especialmente útil en interacción humano-robot (HRI), permitiendo evaluar respuestas de usuarios ante comportamientos de robots antes de su implementación automática completa (Riek, 2012). Esta técnica, en la cual un operador humano controla discretamente las acciones del robot de forma transparente para el usuario, facilita la recopilación de retroalimentación durante las fases tempranas del desarrollo. No obstante, las implementaciones tradicionales de WoZ en robótica social frecuentemente carecen de infraestructuras estandarizadas y modulares que faciliten la participación de implicados no técnicos en el proceso de refinamiento de funcionalidades (Hoffman, 2016; Porfirio et al., 2018).

En respuesta a los desafíos identificados en trabajos previos (Cubero et al., 2024), el presente artículo describe una evolución significativa de la plataforma de evaluación de robots asistenciales. Esta implementación inicial demostró la viabilidad del enfoque de simulación virtual para evaluación de robots sociales. Sin embargo, se detectaron limitaciones arquitectónicas para este tipo de soluciones que son abordadas en este artículo.

Aunque el trabajo presenta una infraestructura generalista, lo ejemplificamos con la versión virtual de un robot social físico existente denominado SHARA (Figura 1) (Villa et al., 2025), un robot conversacional con capacidades afectivas y proactivas, enfocado a la asistencia y acompañamiento de personas mayores que viven solas en su hogar. SHARA cuenta con la capacidad de detectar la presencia de sus usuarios, reconocerlos e iniciar conversaciones de forma proactiva con ellos (Villa et al., 2022). Además, puede contextualizar cada una de sus respuestas gracias a recordar las conversaciones previas (Villa et al., 2022).

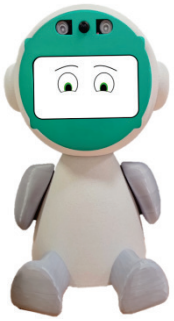


*Figura 1: Robot asistencial social SHARA*

El sistema presentado en este artículo proporciona una simulación virtual del robot asistencial SHARA (imitando gran parte de sus capacidades) y permite el control remoto del robot (virtual en este caso) por parte de un operador y en general realizar evaluaciones con diferentes implicados de forma más flexible.

Las contribuciones principales de este trabajo se centran en un modelo de arquitectura para sistemas basados en la técnica de WoZ para SARs y se abordan sistemáticamente las limitaciones identificadas en trabajos anteriores, materializadas mediante: (1) migración de una aplicación de escritorio desarrollada en PyQt6 que optimiza el rendimiento y reduce significativamente la latencia de comunicación mediante una arquitectura nativa; (2) integración de modelos de lenguaje de OpenAI (GPT-4o-mini1) para generación automática de respuestas contextualizadas que mantienen coherencia conversacional y personalización, además de ser el modelo de conversación que usa el robot real; (3) implementación de un sistema de generación de respuestas múltiples basadas en emociones que proporcionan al operador alternativas adaptadas a un contexto afectivo del robot; (4) rediseño completo de la interfaz de usuario con componentes modulares que facilitan la supervisión simultánea de múltiples canales de interacción

[1]https://platform.openai.com/docs/models/gpt-4o-mini Último acceso: 28/10/2025

(conversación, vídeo, estado de la simulación); y (5) implementación de una arquitectura de sistema orientada a servicios que garantiza escalabilidad, mantenibilidad y futura extensibilidad del sistema.

El presente artículo se estructura de la siguiente manera: 2. Estado del Arte contextualiza las contribuciones en el marco de investigación previas en robótica social asistencial, metodologías WoZ y plataformas de evaluación; 3. Materiales y Métodos describe la arquitectura del sistema mejorado, las tecnologías empleadas y las decisiones de diseño fundamentales; 4. Resultados presenta análisis comparativos detallados entre ambas versiones del sistema, incluyendo mejoras cuantificables en latencia y análisis de capacidades funcionales ampliadas; finalmente, 5. Conclusiones sintetiza los aportes del trabajo y delinean direcciones futuras de investigación en el desarrollo de plataformas WoZ para robótica social asistencial.

## 2. Estado del arte

El desarrollo y evaluación de robots sociales asistenciales representa un campo en constante evolución que combina múltiples dominios de investigación, incluyendo la HRI, tecnologías asistenciales y metodologías de evaluación basadas en simulación. Esta sección proporciona una revisión de los enfoques actuales en la robótica social para el cuidado de personas mayores, identificando los desafíos principales que el sistema propuesto aborda y contextualizando las mejoras implementadas respecto al trabajo previo.

### 2.1 Robots sociales asistenciales para el cuidado de personas mayores

Los robots sociales asistenciales han emergido como un paradigma distintivo en el cual proporcionan asistencia a través de interacción social más que física (Feil-Seifer & Matarić, 2005). Este enfoque ha ganado relevancia significativa para abordar las necesidades crecientes de una población global envejecida, con aplicaciones que abarcan desde recordatorios de medicación hasta acompañamiento social y estimulación cognitiva.

Se identifican revisiones sistemáticas de implementaciones SARs han revelado tanto beneficios potenciales significativos como limitaciones actuales en entornos de cuidado geriátrico (Zhao et al, 2023). Algunos de los factores clave que influyen en la aceptación de los robots son la importancia de interacciones personalizadas, sensibilidad cultural y la necesidad de introducción gradual de dicha tecnología en la vida de sus usuarios (Rigaud et al., 2024).

Para tratar esto factores, varios estudios recientes enfatizan en que el diseño de robots sociales efectivos requiere comprensión profunda de principios de experiencia de usuario y metodologías de diseño centradas en él. El diseño de estos sistemas debe entender la experiencia de usuario en interacción humano máquina, identificar factores clave que influyen en la aceptación y validación del usuario, recalcando que las funcionalidades técnicas por sí solas no son suficientes. Por ello, se deben proporcionar experiencias emocionales positivas en robots sociales y mantener la confianza del usuario durante periodos de interacción extendidos (Alenjung et al., 2019). Al evaluar estos sistemas, debe prestarse especial atención al impacto de la edad en su aceptación y usabilidad (Bevilacqua et al., 2022).

Estas evaluaciones se han destacado por varias investigaciones, enfatizando la importancia de factores culturales y contextuales en la evaluación de robots asistenciales (Papadopoulos et al., 2022). Para ello, las metodologías de evaluación deben considerar diversos antecedentes de usuarios, entornos de cuidado y contextos sociales para asegurar que los resultados de estas evaluaciones sean generalizables a través de diferentes escenarios de implementación (Asi et al., 2022).

### 2.2 Metodologías de evaluación para robótica social

La evaluación de robots sociales para el cuidado geriátrico presenta desafíos únicos que se extienden más allá de las pruebas tradicionales de usabilidad. Los enfoques actuales frecuentemente carecen de la escalabilidad, accesibilidad y rigor metodológico necesarios para realizar evaluaciones en diversos entornos y poblaciones de usuarios (Koh et al., 2021). Teniendo en cuenta que para medir resultados sociales, emocionales y terapéuticos se requieren marcos de evaluación que permitan a varios interesados, de distintas ramas de estudios, evaluar dichos resultados manteniendo validez científica.

Por ello, los enfoques de evaluación de métodos mixtos son particularmente valiosos para capturar la naturaleza multifacética de la efectividad de robots sociales (Asl et al., 2022). Estos enfoques combinan medidas cuantitativas con conocimientos cualitativos, permitiendo a investigadores evaluar tanto resultados medibles como experiencias subjetivas de usuarios, las cuales son muy importantes para entender la aceptación de robots en entornos de cuidado geriátrico. La integración de evaluaciones cuantitativas estructuradas con técnicas de observación etnográfica aporta comprensión de cómo funcionan los robots sociales dentro de contextos de cuidado complejos.

Otra metodología de interés es la evaluación longitudinal para robots sociales en entornos de cuidado geriátrico. A diferencia de estudios de laboratorio de corto plazo, los

enfoques longitudinales rastrean la aceptación del usuario, cambios comportamentales y resultados terapéuticos durante períodos extendidos. Esta metodología es particularmente útil para poblaciones amplias, donde los efectos iniciales de novedad pueden disminuir y emergen patrones verdaderos de aceptación del robot y eficacia terapéutica (Coronado et al., 2022).

De forma similar, las metodologías de evaluación basadas en escenarios proporcionan enfoques estructurados para probar robots sociales a través de diversos contextos de interacción relevantes para el cuidado geriátrico (Koh et al., 2021). Estas metodologías implican diseñar escenarios de cuidado realistas que reflejan desafíos e interacciones comunes en entornos de cuidado geriátrico. Es importante destacar que las metodologías de evaluación participativa involucran activamente a los usuarios mayores, cuidadores y profesionales de salud en el proceso de evaluación, asegurando que los criterios de evaluación reflejen las necesidades y preferencias de los implicados (Tobis et al., 2023).

### 2.3 La técnica Wizard of Oz en robótica social

Entre las metodologías experimentales controladas, la técnica Wizard of Oz (WoZ) representa una metodología de investigación en interacción humano-máquina que ha sido extensivamente adaptada para estudios de interacción humano-robot (HRI) (Riek, 2012). En experimentos WoZ, los participantes interactúan con lo que ellos creen ser un sistema robótico autónomo, mientras que en realidad un operador humano controla de forma encubierta algunas o todas las acciones del robot.

Estos estudios WoZ pueden evaluar efectivamente estrategias de interacción social para robots operando bajo restricciones perceptuales (Sequeira et al., 2016). Incluso cuando los robots tienen capacidades sensoriales limitadas, los estudios WoZ cuidadosamente diseñados pueden revelar patrones de interacción exitosos para posteriormente ser incorporados en sistemas autónomos.

Gracias a estos estudios, WoZ se ha establecido como metodología relevante en investigación HRI por varias razones. En primer lugar, permite prototipados rápidos y pruebas de hipótesis de comportamientos robóticos sin la inversión sustancial de tiempo y recursos requerida para desarrollar sistemas completamente autónomos. En segundo lugar, la metodología proporciona control experimental preciso sobre variables de interacción mientras mantiene el foco en respuestas humanas naturales, lo cual es particularmente valioso cuando se estudian resultados sociales, emocionales y terapéuticos complejos. En tercer lugar, WoZ facilita enfoques de diseño centrados en el usuario, ya que permite a investigadores probar y refinar comportamientos de robots basándose en retroalimentación auténtica de usuarios en entornos realistas antes de comprometerse con la implementación autónoma completa (Porfirio et al., 2018).

A pesar de los avances en robótica social y metodologías de evaluación, aún persisten brechas en sistemas WoZ actuales. Frecuentemente carecen de infraestructuras estandarizadas y modulares que permitirían a los implicados no técnicos participar de forma efectiva en procesos de desarrollo y evaluación de robots (Porfirio et al., 2018). Esta falta de estandarización se manifiesta en las implementaciones ad-hoc desarrolladas con poco tiempo, que suelen sufrir limitaciones como retrasos de entrada, respuestas inconsistentes del operador y falta de protocolos establecidos. Estas deficiencias pueden afectar negativamente tanto la autenticidad del comportamiento del robot como la percepción del participante.

Aunque existen esfuerzos por abordar estos problemas de estandarización, como la plataforma de código abierto WoZ4U, que proporciona atajos de teclado, cadenas de comportamiento automatizadas y acciones del operador optimizadas para reducir la carga cognitiva (Hoffman, 2016), la mayoría de las plataformas existentes se enfocan en sistemas completamente automatizados o controlados por un operador. Este enfoque binario limita su aplicabilidad en escenarios complejos de robótica social.

Para superar esta limitación, los sistemas WoZ híbridos emergen como una solución prometedora. Estos sistemas combinan control del operador con componentes autónomos, representando un avance sobre metodologías tradicionales (Hoffman, 2016). El enfoque híbrido permite evaluaciones más realistas de robots sociales semi-autónomos y facilita transiciones graduales de control del operador hacia autonomía completa, resultando práctico para desarrollar y probar comportamientos adaptativos que deben responder a situaciones sociales impredecibles.

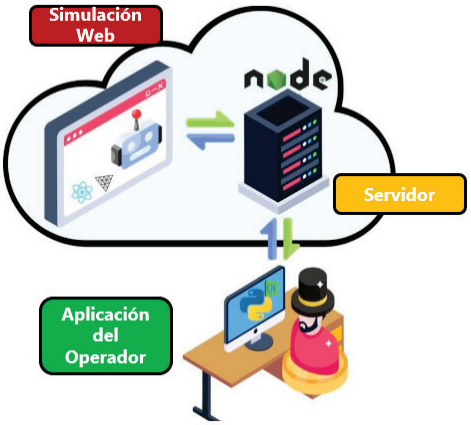



*Figura 2: Arquitectura tripartita del sistema de evaluación remota*

### 2.4 Lecciones aprendidas del Estado del Arte

Gracias a un estudio profundo de la robótica social y los métodos de evaluación más utilizados en la actualidad, podemos ver cómo deberían desarrollarse y evaluarse estos sistemas enfocados a adultos mayores.

En primer lugar, el desarrollo técnico no es lo más importante. Para que un robot sea aceptado, debe generar confianza, ser agradable y entender las limitaciones del contexto cultural de la persona mayor. Un robot útil es aquel que se adapta al usuario y a su entorno.

En segundo lugar, no es suficiente con evaluar el comportamiento y rendimiento del robot en un entorno de laboratorio controlado. Para conocer su verdadero impacto, se necesitan métodos que recojan experiencias reales y estudios que revelen el uso continuo del robot. La evaluación se debe hacer en un entorno realista.

Finalmente, la técnica WoZ híbrida resulta ser la solución ideal para poner en práctica estas lecciones ya que:

- Permite probar ideas rápidamente eliminando la necesidad de programar un robot completamente autónomo. Un operador puede simular comportamientos sociales complejos y ver cómo reaccionan los adultos mayores, refinando el diseño durante el desarrollo.
- El operador humano aporta el sentido común y la empatía para manejar situaciones impredecibles, mientras que el sistema automatizado se encarga de tareas repetitivas. Esto garantiza interacciones fluidas y realistas.
- Permite pasar de manera gradual y controlada de un prototipo controlado por un operador a un robot cada vez más autónomo, incorporando sólo los comportamientos que han sido previamente validados.

La técnica WoZ híbrida es una herramienta que permite

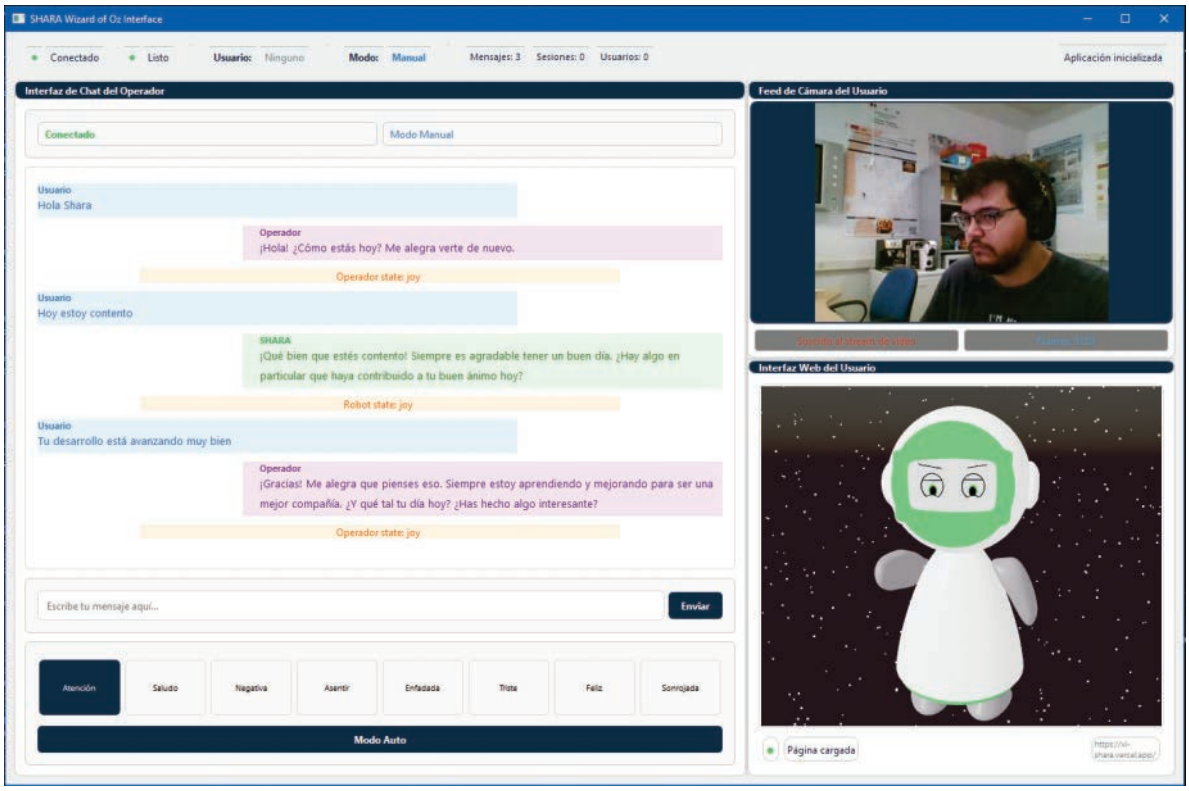



*Figura 3: Interfaz del operador. La interfaz conjunta un historial de chat conversacional de la interacción, vídeo en tiempo real del usuario, navegador con la interfaz web actualizada que ve el usuario y panel de control para una manipulación fuera del flujo conversacional.*

centrarse en los usuarios a la hora de desarrollar robots sociales, puesto que con ella se pueden desarrollar y evaluar interacciones sociales significativas de una manera directa y efectiva.

## 3. Materiales y métodos

Esta sección describe la evolución de la arquitectura del sistema SHARA-WoZ desde su versión original (v1.0.0) hasta la versión mejorada actual (v2.0.0). Primero se presenta el sistema inicial con las limitaciones identificadas mediante evaluación por panel de expertos, que motivaron la reestructuración del sistema.

Posteriormente se detalla la arquitectura del sistema en su siguiente versión, describiendo sus componentes principales, las novedades implementadas y la justificación técnica de las decisiones de diseño que abordan las limitaciones previamente identificadas, manteniendo el mismo modelo de comunicación de componentes.

### 3.1 Sistema inicial (v1.0.0): Arquitectura y limitaciones identificadas

El sistema original SHARA-WoZ v1.0.0 implementaba una arquitectura web tripartita (Figura 2): una aplicación web React/Three.js para simulación 3D del robot, un servidor central Node.js para coordinación de comunicaciones, y una aplicación desktop Python para el operador WoZ (Cubero et al., 2024). Esta arquitectura permitió evaluaciones remotas de interacciones usuario-robot eliminando barreras logísticas del despliegue físico.

La evaluación mediante panel de expertos identificó tres limitaciones críticas que motivaron la re-arquitectura completa. Entre las limitaciones clave, destacan (1) latencia del sistema que comprometía la naturalidad de las interacciones, factor crítico en asistencia geriátrica; (2) ausencia de respuestas pre-generadas que obligaba a redacción manual de cada respuesta del robot; y (3) limitaciones de expresividad emocional. Además, se detectaron problemas técnicos en detección de silencios que provocaban pérdida de mensajes del usuario.
Estas limitaciones evidenciaron la necesidad de priorizar optimización de rendimiento, automatización inteligente de respuestas con preservación de control humano, y modelado emocional sofisticado.

### 3.2 Sistema mejorado (v2.0.0): Arquitectura y componentes

El sistema SHARA-WoZ v2.0.0 implementa una arquitectura modular rediseñada que aborda sistemáticamente las limitaciones identificadas, manteniendo el modelo tripartito de la figura 2, pero con mejoras sustanciales en cada componente.

#### *3.2.1 Aplicación de escritorio del operador*

La interfaz del operador fue completamente reimplementada utilizando PyQt6, *framework* nativo que proporciona mayor rendimiento y capacidades de integración que las tecnologías web empleadas en la primera versión. El sistema se modela organizando sus funcionalidades en servicios especializados que gestionan aspectos específicos del sistema. Estos pueden ser identificados como funciones clave de todo sistema de Mago de Oz para SARs:

- **Gestión de eventos:** Coordina la comunicación entre componentes mediante un sistema de publicación-suscripción desacoplado.
- **Gestión de estado:** Mantiene el estado global del sistema incluyendo conexión, modo de operación y usuario activo.
- **Comunicación en tiempo real:** Maneja la conexión bidireccional con el servidor mediante sockets con reconexión automática.
- **Procesamiento de mensajes:** Coordina el flujo conversacional entre modos automático y manual.
- ***Streaming* de vídeo:** Procesa y renderiza el *feed* de video del usuario en tiempo real.

La implementación particular de la interfaz presenta tres paneles principales (Figura 3): un panel de chat con historial conversacional y controles de estado emocional del robot, un panel de vídeo que muestra al usuario en tiempo real, y un navegador web integrado que replica la interfaz que ve el usuario final. Se incluye un panel de control inferior que permite la manipulación de la simulación fuera del flujo conversacional habitual, de forma que el operador puede manipular la simulación a todos los niveles.

#### *3.2.2 Integración con inteligencia artificial*

La integración con servicios de OpenAI representa la innovación más significativa del sistema en esta versión. La propuesta arquitectónica aprovecha los avances de la IA generativa y los sistemas basados en LLMs para una generación dinámica de posibles respuestas del robot. El servidor implementa dos estrategias complementarias de generación de respuestas:

- **Generación de respuesta principal:** El sistema genera una respuesta conversacional utilizando un contexto temporal e identificación de usuario para mayor personalización en las conversaciones, de la misma forma que el robot físico.
- **Generación de variantes emocionales:** La innovación consiste en generar simultáneamente ocho variantes de la respuesta principal, cada una adaptada a un estado

emocional específico del robot (sorpresa, neutral, negación, afirmación, enfado, tristeza, alegría, sonrojada). Las variantes se generan en paralelo para minimizar latencia, proporcionando al operador opciones contextualizadas listas para envío inmediato.

Esta arquitectura de respuestas múltiples reduce en gran medida el tiempo de respuesta en modo semi-automático mientras que preserva la capacidad de supervisión y modificación humana (Figura 4).

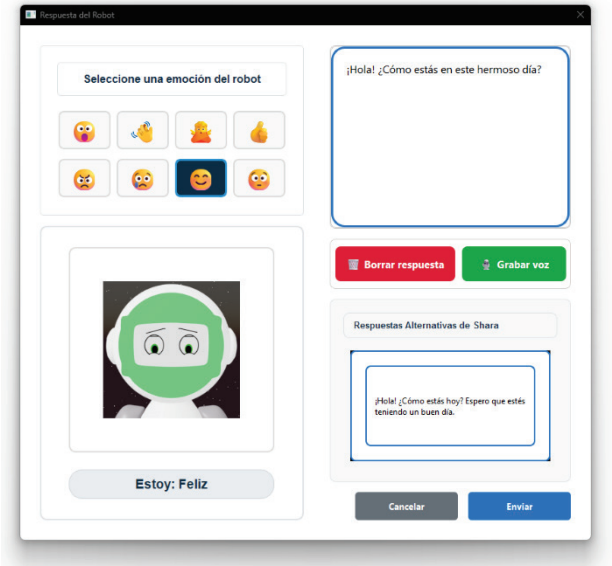



*Figura 4: Interfaz de respuesta del operador. Muestra la respuesta generada por la simulación y para cada uno de sus estados aporta una respuesta alternativa. El operador puede modificar las respuestas o aceptarlas.*

### *3.2.3 Servidor central y servicios cloud*

Se dispone de un servidor Node.js como centro de comunicación, que incorpora optimizaciones significativas. Gestiona conversaciones por usuario con persistencia de historial completo, coordina el envío diferencial de mensajes según el modo operacional activo (automático, semi-automático o manual), e integra servicios cloud de Google para transcripción de audio2 y síntesis de voz3 con configuración optimizada como el robot físico.

El servidor implementa lógica de coordinación que en modo automático envía respuestas directamente sin intervención del operador, en modo semi-automático presenta opciones al operador para aprobación/modificación, y en modo manual permite composición completa por el operador.

[2] https://cloud.google.com/speech-to-text Último acceso: 04/11/2025

[3] https://cloud.google.com/text-to-speech Último acceso: 04/11/2025

### *3.2.4 Interfaz web del usuario final*

La interfaz React/Three.js del usuario final se encarga de la simulación 3D validada del robot con animaciones faciales y corporales sincronizadas con estados emocionales. Esta continuidad garantiza comparabilidad con evaluaciones previas (Figura 5).

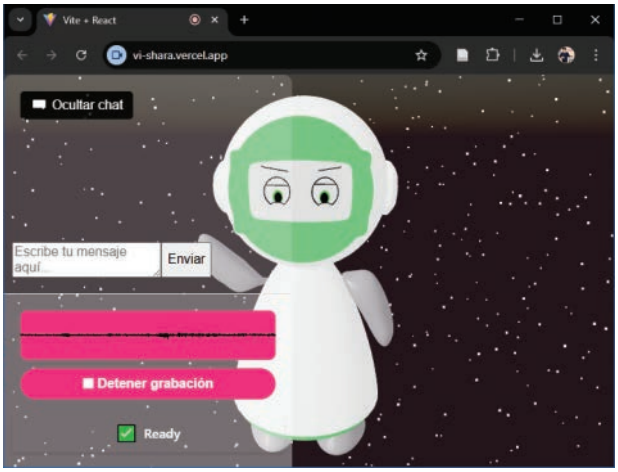



*Figura 5: Interfaz web de la simulación del robot SHARA. Muestra una representación 3D y da la opción de ver un historial conversacional en forma de chat al igual que un indicador del estado de la simulación.*

## 3.3 Justificación de decisiones de diseño

Las decisiones arquitectónicas responden sistemáticamente a las limitaciones identificadas y constituyen las contribuciones principales del artículo al aportar un modelo arquitectónico y basado en evidencias de los sistemas de Mago de Oz para SARs:

**Migración a PyQt6 nativo:** Reduce latencia mediante eliminación de sobrecarga de comunicación navegador-aplicación y proporciona procesamiento multi-hilo eficiente para vídeo y mensajes sin bloqueo de interfaz. Las evaluaciones internas mostraron una reducción del 40% en tiempo de renderizado comparando con la implementación web híbrida de la v1.0.0.

**Sustitución de servicio de generación de respuestas:** El sistema previo usaba la combinación de los servicios de Watson Assistant, Natural Language Understanding (NLU) y los servicios de traducción de Google para generar las respuestas del robot, mientras que la nueva implementación usa los servicios de GPT como sustitución, ya que elimina la necesidad de traducción y engloba las dos funcionalidades que se empleaban con los servicios de Watson.

**Sistema de respuestas múltiples:** Aborda simultáneamente las limitaciones de "respuestas pre-generadas" y "protocolo de comunicación social" proporcionando opciones

contextualizadas que mantienen al operador en el bucle de decisión para juicios clínicos sobre apropiación emocional.

**Arquitectura modular:** Facilita extensibilidad para incorporación futura de funcionalidades identificadas por expertos (detección de emergencias, análisis de comportamiento no verbal).

Esta arquitectura establece una base extensible para mejoras futuras mientras resuelve las limitaciones identificadas por expertos en el sistema original.

## 4. Resultados

Esta sección presenta una evaluación comparativa entre las versiones v1.0.0 y v2.0.0 del sistema SHARA-WoZ, analizando métricas de rendimiento para aplicaciones de robótica social asistencial. Las mediciones se realizaron en condiciones controladas para garantizar comparabilidad entre ambas arquitecturas.

### 4.1 Métricas de latencia conversacional

La latencia de respuesta conversacional constituye un factor crítico en interacciones usuario-robot, especialmente en contextos geriátricos donde la naturalidad de la comunicación impacta directamente en la aceptación del sistema (Wang, Y. L., & Lo, C. W. 2025). La tabla 1 presenta una comparativa detallada de los tiempos de respuesta entre los servicios de procesamiento de lenguaje natural empleados en cada versión.

*Tabla 1: Comparativa de latencia en servicios de procesamiento conversacional*

| Componente | V1.0.0 | V2.0.0 | Reducción |
|---|---|---|---|
| Transcripción de voz | Google STT: ~1000 ms | Google STT: ~1000 ms | 0% |
| Traducción | Google Translate: 800-1300 ms | No requerido | 100% |
| Síntesis de voz | Google TTS: 500-800 ms | Google TTS: 500-800 ms | 0% |
| Procesamiento IA | IBM Watson Assistant + IBM NLU: 1500-3000 ms | GPT-4o-mini: 200-500 ms | 80-85% |
| Latencia total (promedio) | 3800-5100 ms | 1700-2300 ms | 55% |
| Latencia p95 | 6500 ms | 3000 ms | 54% |

La implementación de GPT-4o-mini representa la mejora más significativa, eliminando la necesidad de servicios de traducción intermedios y proporcionando respuestas con latencias optimizadas para aplicaciones interactivas en tiempo real. La generación paralela de respuestas emocionales mantiene tiempos competitivos mediante procesamiento concurrente.

### 4.2 Rendimiento de interfaz del operador

La migración de una arquitectura híbrida web-nativa a PyQt6 puro impacta significativamente en la velocidad de respuesta de la interfaz del operador, como se detalla en la tabla 2.

*Tabla 2: Métricas de rendimiento de interfaz del operador*

| Métrica | V1.0.0 | V2.0.0 | Mejora |
|---|---|---|---|
| Tiempo de inicialización | 3000-5000 ms | 800-1200 ms | 73% |
| Renderización inicial UI | 450-600 ms | 250-350 ms | 42% |
| Actualización de chat | 120-180 ms | 50-80 ms | 56% |
| Procesamiento de vídeo (*frame*) | 45-60 ms | 25-35 ms | 42% |
| Uso de memoria (base) | 280-350 MB | 120-180 MB | 49% |
| Uso de CPU (*Idle*) | 8-12% | 3-5% | 58% |

La eliminación de la sobrecarga del renderizado web y la utilización de bibliotecas nativas de Qt resulta en mejoras sustanciales de rendimiento.

### 4.3 Comparativa de arquitectura de despliegue

Por último, la tabla 3 analiza las diferencias en latencia y disponibilidad entre el despliegue local de v1.0.0, para el cual se plantea un despliegue en un servidor local, y la infraestructura cloud de v2.0.0.

*Tabla 3: Métricas de infraestructura y despliegue*

| Aspecto | V1.0.0 (Local) | V2.0.0 (Cloud) | Diferencia |
|---|---|---|---|
| Latencia red | <5 ms (LAN) | 15-30 ms | +10-25 ms |
| Disponibilidad | 95-98% | 99.9% | +2-4% |
| Tiempo de despliegue | 15-30 min | 2-5 min | 85% reducción |
| Escalabilidad | Manual | Automática | - |
| *Cold start* (primera solicitud) | 0 ms | ~500 ms | +~500 ms |
| Coste/mes | >0€ (servidor local) | 0€ (plan gratuito) | Variable |

Aunque la arquitectura *cloud* introduce latencias de red adicionales, estas son compensadas por las mejoras en el procesamiento de IA y la disponibilidad global. Los servicios *edge* de Vercel minimizan la latencia para usuarios

geográficamente distribuidos mediante caching y distribución global.

## 5. Conclusiones

El presente trabajo ha descrito el desarrollo e implementación del sistema SHARA-WoZ v2.0.0, una evolución de la plataforma de evaluación para robots sociales asistenciales que aborda sistemáticamente las limitaciones identificadas en su versión anterior. Los resultados obtenidos validan las decisiones arquitectónicas adoptadas y demuestran mejoras cuantificables en múltiples dimensiones críticas para la evaluación efectiva de interacciones humano-robot para adultos mayores.

### 5.1 Lecciones aprendidas arquitecturales

Es posible identificar cinco aspectos fundamentales en sistemas de Mago de Oz para SARs que se han validado exitosamente a través de métricas objetivas. La migración a PyQt6 nativo ha resultado en una reducción del 42% en tiempos de renderizado y del 49% en uso de memoria, confirmando las ventajas de arquitecturas nativas sobre soluciones híbridas web. La integración de GPT-4o-mini ha transformado el sistema conversacional, reduciendo la latencia total en un 55% y eliminando la necesidad de servicios de traducción intermedios. El sistema de generación de respuestas múltiples por estado emocional ha reducido el tiempo de respuesta en modo semi-automático de 15-30 segundos a 2-3 segundos, representando una mejora del 87-93% mientras preserva el control humano esencial en aplicaciones clínicas.

La arquitectura modular implementada facilita la extensibilidad del sistema sin refactorización mayor, como evidencian los servicios especializados desacoplados que gestionan eventos, estado, comunicación y procesamiento de manera independiente. El rediseño de la interfaz del operador, basado en patrones de diseño modernos, ha mejorado significativamente la supervisión simultánea de múltiples canales de interacción.

### 5.2 Implicaciones para la robótica social asistencial

Este trabajo contribuye al campo emergente de la robótica social asistencial proporcionando una plataforma de evaluación que equilibra sofisticación técnica con aplicabilidad práctica. La experiencia acumulada durante el desarrollo del sistema permite identificar aspectos fundamentales que todo sistema de Mago de Oz para robots sociales asistenciales debería contemplar.

En primer lugar, la minimización de la latencia conversacional emerge como factor clave. Los resultados demuestran que latencias superiores a 3 segundos comprometen significativamente la naturalidad de las interacciones, especialmente en contextos geriátricos donde la fluidez comunicativa impacta directamente en la aceptación del sistema. La arquitectura debe priorizar la optimización de cada componente del flujo conversacional, desde la transcripción de voz hasta la síntesis de respuestas.

En segundo lugar, la gestión modular de funcionalidades mediante servicios especializados facilita tanto el desarrollo como el mantenimiento del sistema. La separación de responsabilidades en servicios independientes para gestión de eventos, estado, comunicación en tiempo real, procesamiento de mensajes y streaming de vídeo permite escalabilidad, extensibilidad y facilita la incorporación de nuevas capacidades sin comprometer la estabilidad del sistema existente.

En tercer lugar, la integración de inteligencia artificial conversacional contextualizada representa un elemento diferenciador. La capacidad de generar respuestas personalizadas considerando el historial conversacional y el contexto temporal del usuario, combinada con la generación simultánea de variantes emocionales, proporciona al operador herramientas para mantener interacciones naturales y coherentes sin sacrificar la supervisión humana en aplicaciones clínicas.

Finalmente, la capacidad del sistema para operar en modos automático, semi-automático y manual responde a la necesidad identificada en la literatura de transiciones graduales desde control humano completo hacia autonomía robótica. Esta flexibilidad operacional resulta esencial en contextos geriátricos donde la adaptabilidad y sensibilidad contextual son cruciales para la aceptación y efectividad terapéutica del sistema.

### 5.3 Trabajo futuro

Las direcciones futuras de desarrollo se centran en tres áreas prioritarias identificadas durante la evaluación del sistema.

La implementación de un repertorio gestual ampliado y expresiones faciales mejoradas aumentará la naturalidad de las interacciones. Se desarrollarán comportamientos proactivos que permitan al robot iniciar conversaciones contextualmente apropiadas, junto con un sistema de memoria conversacional a largo plazo para mantener continuidad entre sesiones.

El rediseño de la interfaz seguirá principios de diseño centrado en la tarea, simplificando componentes y facilitando el control de la simulación y de la propia interfaz del operador.

Estas mejoras se implementarán iterativamente con validación continua por usuarios finales y expertos del

dominio, manteniendo el equilibrio entre funcionalidad avanzada y simplicidad operacional que caracteriza al sistema actual.

Además, se plantea la evaluación y validación de este sistema con usuarios expertos en el ámbito de la interacción persona ordenador, persona robot y expertos en el ámbito sanitario para evaluar la calidad conversacional de la simulación. La evaluación se centraría en evaluar los componentes de la interfaz en su usabilidad, y cómo de sencillo resultaría emplear esta forma de evaluar robots sociales durante su desarrollo.

### 5.4 Contribución

El sistema SHARA-WoZ v2.0.0 representa un avance significativo en la evolución de plataformas de evaluación para robótica social asistencial. Al abordar sistemáticamente las limitaciones técnicas mientras mantiene la flexibilidad operacional, este trabajo establece un precedente para el desarrollo de herramientas de evaluación que equilibran rigurosidad técnica con aplicabilidad práctica. La naturaleza modular y extensible de la arquitectura propuesta facilita su adaptación a diversos contextos de investigación y dominios de aplicación más allá del cuidado geriátrico.

Los resultados demuestran que es posible desarrollar sistemas WoZ que proporcionen tanto la sofisticación técnica requerida por investigadores como la accesibilidad necesaria para implicados técnicos. Esta democratización del acceso a herramientas de evaluación robótica constituye un paso importante hacia la adopción de robots sociales en contextos asistenciales, contribuyendo en última instancia a mejorar la calidad de vida de poblaciones vulnerables mediante tecnología centrada en el humano.

## Agradecimientos

Esta investigación fue financiada por la JUNTA DE COMUNIDADES DE CASTILLA-LA MANCHA mediante las ayudas números SBPLY/21/180501/000160 (SHARA3) y SBPLY/24/180225/000176 (AKAI-SHARA); y por el MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES mediante la ayuda número FPU22/00839 (contrato predoctoral).

## Referencias

Abdi, J., Al-Hindawi, A., Ng, T., & Vizcaychipi, M. P. (2018). Scoping review on the use of socially assistive robot technology in elderly care. *BMJ Open*, *8*(2), e018815. https://doi.org/10.1136/bmjopen-2017-018815

Alenjung, B., Lindblom, J., Andreasson, R., & Ziemke, T. (2019). User experience in social human-robot interaction. In *Rapid automation: Concepts, methodologies, tools, and applications* (pp. 1468-1490). IGI Global.

Asi, A. M., Ulate, M. M., Martin, M. F., & van der Roest, H. (2022). Methodologies used to study the feasibility, usability, efficacy, and effectiveness of social robots for elderly adults: Scoping review. *Journal of Medical Internet Research*, *24*(8), e37434. https://doi.org/10.2196/37434

Bevilacqua, R., Di Rosa, M., Riccardi, G. R., Pelliccioni, G., Lattanzio, F., Felici, E., Margaritini, A., Amabili, G., & Maranesi, E. (2022). Design and development of a scale for evaluating the acceptance of social robotics for older people: The robot era inventory. *Frontiers in Neurorobotics*, *16*, 883106. https://doi.org/10.3389/fnbot.2022.883106

Coronado, E., Kiyokawa, T., Garcia Ricardez, G. A., Ramirez-Alpizar, I. G., Venture, G., & Yamanobe, N. (2022). Evaluating quality in human-robot interaction: A systematic search and classification of performance and human-centered factors, measures and metrics towards an industry 5.0. *Journal of Manufacturing Systems*, *63*, 392-410. https://doi.org/10.1016/j.jmsy.2022.04.007

Cubero, G., Villa, L., Favela, J., Ochoa, S., Díaz-Fernández, C., & Hervás, R. (2024, November). SHARA in the Land of Oz: A Platform for the Rapid Development of Human-Robot-Interactions Using Social Robot Simulator and Wizard of Oz. In *International Conference on Ubiquitous Computing and Ambient Intelligence* (pp. 87-98). Cham: Springer Nature Switzerland. Feil-Seifer, D., & Matarić, M. J. (2005). Defining socially assistive robotics. In *Proceedings of the 2005 IEEE 9th International Conference on Rehabilitation Robotics* (pp. 465-468). IEEE. https://doi.org/10.1109/ICORR.2005.1501143

Hoffman, G. (2016). OpenWoz: A runtime-configurable wizard-of-oz framework for human-robot interaction. In *AAAI Spring Symposium - Technical Report* (Vol. SS-16-01-07, pp. 121-126).

Kachouie, R., Sedighadeli, S., Khosla, R., & Chu, M. T. (2014). Socially assistive robots in elderly care: A mixed-method systematic literature review. *International Journal of Human-Computer Interaction*, *30*(5), 369-393. https://doi.org/10.1080/10447318.2013.873278

Koh, W. Q., Felding, S. A., Budak, K. B., Toomey, E., & Casey, D. (2021). Barriers and facilitators to the implementation of social robots for older adults and people with dementia: A scoping review. *BMC Geriatrics*, *21*(1), 351. https://doi.org/10.1186/s12877-021-02277-9

Porfirio, D., Sauppé, A., Albarghouthi, A., & Mutlu, B. (2018). Authoring and verifying human-robot interactions. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (pp. 75-86). https://doi.org/10.1145/3242587.3242634

Riek, L. D. (2012). Wizard of Oz studies in HRI: A systematic review and new reporting guidelines. *Journal of Human-Robot Interaction*, *1*(1), 119-136. https://doi.org/10.5898/JHRI.1.1.Riek

Rigaud, A. S., Dacunha, S., Harzo, C., Lenoir, H., Sfeir, I., Piccoli, M., & Pino, M. (2024). Implementation of socially assistive robots in geriatric care institutions: Healthcare professionals' perspectives and identification of facilitating factors and barriers. *Journal of Rehabilitation and Assistive Technologies Engineering*, *11*, 20556683241284765.

Sequeira, P., Alves-Oliveira, P., Ribeiro, T., Di Tullio, E., Petisca, S., Melo, F. S., Castellano, G., & Paiva, A. (2016). Discovering social interaction strategies for robots from restricted-perception wizard-of-oz studies. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 197-204). IEEE. https://doi.org/10.1109/HRI.2016.7451752

Tobis, S., Piasek-Skupna, J., Neumann-Podczaska, A., Suwalska, A., & Wieczorowska-Tobis, K. (2023). The effects of stakeholder perceptions on the use of humanoid robots in care for older adults: Post-interaction cross-sectional study. *Journal of Medical Internet Research*, *25*, e46617. https://doi.org/10.2196/46617

Villa, L., Hervás, R., Cabañero, L., Fontecha, J., López, G., & Favela, J. (2025). Evaluating proactivity levels in socially assistive robots for elderly care: a user adoption assessment. *Behaviour & Information Technology*, 1-26.

Villa, L., Hervás, R., Dobrescu, C. C., Cruz-Sandoval, D., & Favela, J. (2022, June). Incorporating affective proactive behavior to a social companion robot for community dwelling older adults. In *International Conference on Human-Computer Interaction* (pp. 568-575). Cham: Springer Nature Switzerland.

Villa, L., Hervás, R., Cruz-Sandoval, D., & Favela, J. (2022, November). Design and evaluation of proactive behavior in conversational assistants: approach with the eva companion robot. In *International Conference on Ubiquitous Computing and Ambient Intelligence* (pp. 234-245). Cham: Springer International Publishing.Zhao, D., Sun, X., Shan, B., Yang, Z., Yang, J., Liu, H., Jiang, Y., & Hiroshi, Y. (2023). Research status of elderly-care robots and safe human-robot interaction methods. *Frontiers in Neuroscience*, *17*, 1291682. https://doi.org/10.3389/fnins.2023.1291682

Papadopoulos, C., Castro, N., Nigath, A., Davidson, R., Faulkes, N., Menicatti, R., ... & Sgorbissa, A. (2022). The CARESSES randomised controlled trial: exploring the health-related impact of culturally competent artificial intelligence embedded into socially assistive robots and tested in older adult care homes. *International Journal of Social Robotics*, *14*(1), 245-256.

Wang, Y. L., & Lo, C. W. (2025). The effects of response time on older and young adults' interaction experience with Chatbot. *BMC psychology*, 13(1), 150.

INTERACCION
Revista Digital de AIPO
Asociación Interacción Persona-Ordenador
Vol. 6, No 2 (2025)
ISSN 2695-6578

# Más allá de las hojas de cálculo: creando flujos para la definición, validación e interoperabilidad de variables clínicas

## Beyond spreadsheets: creating workflows for the definition, validation, and interoperability of clinical variables

**Andrea Vázquez-Ingelmo**
Grupo de investigación GRIAL, Departamento de Informática y Automática
Universidad de Salamanca
Salamanca, España
andreavazquez@usal.es

**Islem Román Nieto-Campo**
Grupo de investigación GRIAL
Universidad de Salamanca
Salamanca, España
islemr@usal.es

**Alicia García-Holgado**
Grupo de investigación GRIAL, Departamento de Informática y Automática
Universidad de Salamanca
Salamanca, España
aliciagh@usal.es

**Francisco José García Peñalvo**
Grupo de investigación GRIAL, Departamento de Informática y Automática
Universidad de Salamanca
Salamanca, España
fgarcia@usal.es

**Antonio Sánchez-Puente**
Departamento de Cardiología, Hospital Universitario de Salamanca, SACyL, IBSAL, Facultad de Medicina (Universidad de Salamanca), CIBERCV (ISCiii)
Salamanca, España
sanchezpu@saludcastillayleon.es

**Pedro L. Sánchez**
Departamento de Cardiología, Hospital Universitario de Salamanca, SACyL, IBSAL, Facultad de Medicina (Universidad de Salamanca), CIBERCV (ISCiii)
Salamanca, España
plsanchez@saludcastillayleon.es

Recibido: 12.11.2025 | Aceptado: 05.12.2025

**Palabras Clave**

Gestión de datos
Variables clínicas
Flujo de interacción
Hojas de cálculo
Diseño centrado en el usuario

**Resumen**

Las hojas de cálculo siguen siendo el estándar para definir y recoger variables clínicas, pero su flexibilidad las vuelve frágiles y propensas a errores. Presentamos un rediseño centrado en la persona del flujo de definición de variables dentro de una plataforma que integra datos clínicos estructurados e imágenes médicas. La propuesta sustituye la creación manual del esquema en procesadores de hojas de cálculo por un editor web interactivo y la generación automática de plantillas validadas a partir del esquema interno de la plataforma, reduciendo la carga cognitiva y previniendo errores de formato y semántica, con retroalimentación accionable en la carga. El flujo BPMN actualizado conecta el modelado de variables con una entrada de datos guiada y validaciones, cumpliendo heurísticas clave como visibilidad del estado, prevención y recuperación de errores. Entre las limitaciones persiste la entrada de datos fuera de línea; como trabajo futuro se plantean estudios de usabilidad, interoperabilidad semántica y asistencia con inteligencia artificial para sugerir variables.

**Keywords**

Data management
Clinical variables
Interaction flow
Spreadsheets
User centered design

**Abstract**

Spreadsheets are still the default for defining and collecting clinical variables, but their flexibility makes them fragile and error-prone. We present a user-centered redesign of the variable-definition workflow inside a platform that manages structured clinical data alongside medical images. The proposal replaces manual, spreadsheet-based schema creation with an interactive web editor and automatically generated, validated spreadsheet templates derived from the platform's internal schema. This shift reduces cognitive load, prevents common formatting and semantic errors, and offers immediate, actionable feedback during data

upload. The updated BPMN workflow connects variable modeling with guided data entry and validation, addressing key usability heuristics such as system status visibility, error prevention, and recovery. Remaining limitations (continued offline data entry) and future directions include usability studies, semantic interoperability, and artificial intelligence-assisted variable suggestions.

## 1. Introducción

En la investigación médica y en los flujos de trabajo clínicos, las hojas de cálculo siguen siendo la herramienta predominante para definir, gestionar e intercambiar datos estructurados debido a su accesibilidad y flexibilidad (Iyengar et al., 2019). No obstante, esas mismas virtudes traen consigo desafíos importantes para la calidad de los datos, su trazabilidad y la interoperabilidad entre equipos y sistemas. Diversos estudios han mostrado que las hojas de cálculo empleadas en entornos sanitarios reales contienen errores críticos con elevada frecuencia: las tasas medias de error por celda superaban el 13 % (Dobell et al., 2018). Estos fallos se asocian con diseños estructurales deficientes, prácticas de uso inconsistentes y la ausencia de mecanismos de validación incorporados.

Nuestras previas experiencias en el desarrollo de plataformas que combinan datos clínicos estructurados e imágenes médicas han puesto de relieve las limitaciones de los flujos impulsados por hojas de cálculo (García-Peñalvo et al., 2021, 2024). Definir variables a partir de plantillas exige a los usuarios atravesar un proceso complejo y multietapa: codificar manualmente metadatos en varias hojas, respetar convenciones estrictas de nomenclatura y tipado, y alinear definiciones entre niveles de paciente, estudio y fichero. El resultado es una tarea costosa y propensa a errores, particularmente para perfiles no técnicos; problemas como identificadores reservados o mal formateados, nombres de variables inconsistentes o formatos incorrectos de fechas y valores suelen desembocar en cargas fallidas y ciclos de corrección que interrumpen el trabajo investigador.

Más allá del esfuerzo, existen carencias de interacción que agravan el problema. Aun contando con plantillas y documentación, las hojas de cálculo carecen de guía interactiva, apoyo contextual y retroalimentación en tiempo real; en la práctica, la responsabilidad de garantizar la integridad de los datos recae en los clínicos e investigadores, cuya pericia rara vez está en la modelización y validación de datos. Esta situación se traduce en barreras de incorporación, dependencia de conocimiento tácito y dificultad para mantener la coherencia semántica a lo largo del tiempo.

Como antecedente directo de este trabajo, desarrollamos CARTIER-IA, una plataforma integrada para gestionar conjuntamente datos clínicos estructurados e imágenes médicas en proyectos colaborativos entre instituciones (García-Peñalvo et al., 2021). Con ella habilitamos la definición de variables en tres niveles (paciente, estudio y fichero) y mecanismos de validación automática y seudonimización para mejorar la calidad y la seguridad de los datos. Sin embargo, la definición del esquema se apoyaba en plantillas de Excel: por cada tabla del proyecto había que crear una hoja cuyo nombre coincidiera exactamente con el de la tabla y codificar los metadatos de cada variable en un esquema multicolumna (cabeceras SV0…SV6 para nombre, tipo, valores permitidos, unidades, etc.). Estas convenciones estrictas y poco autoexplicativas, junto con la falta de guía en contexto o vista previa dentro de la propia plataforma, elevaban la probabilidad de error (identificadores inválidos, tipos incorrectos, incoherencias) y complicaban la evolución de los esquemas en estudios multicéntricos o longitudinales, donde las definiciones cambian con frecuencia.

Este trabajo presenta una propuesta que replantea la definición de variables como parte de una plataforma reforzada para gestión de datos e imágenes. La solución sitúa la interacción centrada en la persona en el núcleo del proceso y guía paso a paso la creación, validación y aplicación de variables estructuradas, integrando conocimiento de dominio y reglas de validación directamente en el flujo. Con ello se busca mejorar la calidad de los datos, facilitar el onboarding y favorecer la colaboración entre perfiles técnicos y clínicos, reduciendo la dependencia de manipulaciones manuales en hojas de cálculo.

En síntesis, partimos de un contexto donde la ubicuidad de las hojas de cálculo convive con tasas elevadas de error, sobrecarga cognitiva y falta de trazabilidad. Frente a ello, proponemos un rediseño que desplaza la complejidad desde el usuario hacia la plataforma, proporciona validaciones tempranas y hace visibles los estados y consecuencias de las acciones. Esta orientación pretende no solo disminuir fallos y retrabajo, sino también sentar las bases para prácticas más sostenibles y reproducibles en la definición de variables clínicas.

El resto del artículo se organiza así: en la sección 2 se revisa el estado del arte y la experiencia previa con la plataforma CARTIER-IA; la sección 3 analiza en detalle los problemas

detectados (rigidez de plantillas, falta de retroalimentación, accesibilidad y escalabilidad); la sección 4 presenta la propuesta de rediseño y sus principios, incluyendo modelado visual asistido y validación en tiempo real; la sección 5 describe la implementación y resultados preliminares, con generación automática de plantillas, carga validada y trazabilidad; la sección 6 discute los hallazgos en relación con heurísticas de usabilidad y abre líneas de extensión como versionado e interoperabilidad semántica; finalmente, la sección 7 resume las conclusiones y plantea las líneas futuras de investigación.

## 2. Antecedentes

La definición de variables estructuradas es un pilar de la gestión de datos clínicos porque determina la validez de la investigación, la interoperabilidad entre sistemas y la calidad de los análisis. Tradicionalmente, este modelado se ha sustentado en diccionarios de datos en hojas de cálculo o en formularios electrónicos relativamente rígidos.

Ambos enfoques requieren pericia técnica para especificar campos, tipos y restricciones, y suelen ser frágiles ante cambios del esquema. En entornos clínicos reales, donde los protocolos evolucionan, los equipos son multidisciplinares y los plazos son ajustados, esta fragilidad se traduce en iteraciones costosas, errores de formato y cargas de control de calidad que recaen sobre perfiles cuyo foco principal no es la modelización de datos. La necesidad es doble: flexibilizar la iteración del modelo y proporcionar más guía y retroalimentación al usuario durante el diseño.

### 2.1 Plataformas de captura electrónica de datos (EDC): avances y límites

Las plataformas EDC han democratizado la recogida de datos clínicos al ofrecer validaciones, trazabilidad y flujos reproducibles. REDCap, por ejemplo, permite definir campos desde una interfaz web o mediante la carga de un diccionario en hoja de cálculo (Harris et al., 2009). Sin embargo, cuando un proyecto pasa a producción, el control de cambios del esquema se restringe, lo que dificulta las iteraciones frecuentes típicas de los estudios reales. Además, suelen faltar funciones de interacción avanzadas orientadas al diseño colaborativo, como retroalimentación contextual, coedición en tiempo real o asistencia durante la configuración de validaciones.

OpenClinica y Castor ilustran enfoques complementarios (plantillas Excel en el primer caso y construcción "arrastrar y soltar" en el segundo), pero persisten retos transversales: escalabilidad del modelo cuando crece el número de variables, costes de licencia o de integración, y fricciones al interoperar con repositorios de terceros, especialmente si el estudio combina datos clínicos con imagen médica. En síntesis, estas herramientas han ampliado el acceso, pero siguen tensionando tres aspectos clave: (i) iteración rápida del modelo, (ii) colaboración multiinstitucional y (iii) guía al usuario durante el diseño.

### 2.2 Gestión de imagen médica: alcance y carencias

En el ámbito de la imagen médica, XNAT es la referencia para almacenar, gestionar y compartir estudios DICOM, incluyendo la extensión de metadatos (Herrick et al., 2016). No obstante, su despliegue y mantenimiento exigen una alta cualificación técnica y su cobertura de "variables clínicas generales" es limitada frente a las EDC. Esta asimetría se vuelve crítica cuando el objetivo es alinear, dentro de un mismo flujo operativo, variables clínicas y datos de imagen con garantías de consistencia, trazabilidad y validación.

### 2.3. Experiencia previa: CARTIER-IA y el uso de plantillas

Para abordar parte de esta brecha, desarrollamos CARTIER-IA, una plataforma que integra datos estructurados e imagen médica (García-Peñalvo et al., 2021). El sistema permitió definir variables en distintos niveles (paciente, estudio y fichero), incorporar mecanismos de validación y seudonimización, y vincular las variables a estudios DICOM. El flujo de configuración se basaba en plantillas Excel con convenciones estrictas (por ejemplo, hojas por tabla y columnas SV0…SV6 para codificar nombre, tipo y atributos), lo que resultaba eficaz para equipos técnicos pero poco tolerante a errores y con barreras de entrada para usuarios no expertos.

En la práctica, los fallos más frecuentes (nombres de hoja imprecisos, tipos mal codificados, dominios de valores inconsistentes) provocaban rechazos de carga y ciclos de corrección manual. A ello se sumaba la ausencia de versionado y reutilización de esquemas entre proyectos, especialmente problemática en contextos longitudinales o multicentro. Aunque CARTIER-IA demostró el valor de unificar datos estructurados e imagen, también evidenció que un flujo centrado en plantillas no resuelve por sí solo las necesidades de iteración ágil, asistencia al usuario y colaboración en tiempo real.

El estado actual muestra dos fortalezas y un cuello de botella. Por un lado, las EDC han profesionalizado la captura clínica y XNAT ha consolidado la gestión de imagen; por otro, la combinación de ambos mundos, con iteración rápida del modelo, validaciones asistidas y alineamiento clínico-imagen en un único flujo, sigue siendo un desafío. Nuestra experiencia con CARTIER-IA confirma la oportunidad de avanzar hacia herramientas que mantengan la robustez de las plataformas existentes pero añadan: (i) edición estructural

guiada, (ii) validaciones en tiempo real y (iii) mecanismos de versionado y reutilización de esquemas para escenarios colaborativos y multicentro.

## 3. Análisis de los problemas de CARTIER-IA

El desarrollo de CARTIER-IA permitió comprobar la viabilidad de una plataforma capaz de integrar, en un mismo entorno, datos estructurados, estudios DICOM y ejecución de modelos de inteligencia artificial. El sistema alcanzó una madurez funcional suficiente para su uso en escenarios reales y demostró que la automatización puede reducir significativamente los tiempos de análisis y mejorar la trazabilidad de los datos clínicos. Sin embargo, la experiencia acumulada durante su despliegue reveló limitaciones estructurales y de interacción que condicionan su escalabilidad y adopción por parte de usuarios no técnicos.

Aunque CARTIER-IA se diseñó desde una filosofía de "flujo unificado", la mayor parte de su configuración dependía de mecanismos manuales basados en plantillas Excel estructuradas, que actuaban como puente entre los usuarios clínicos y la base de datos relacional interna. Esta decisión inicial, motivada por la facilidad de edición y la compatibilidad con herramientas de uso común, acabó convirtiéndose en una fuente de rigidez, fragilidad y errores recurrentes. Como se muestra en el flujo de trabajo BPMN (Figura 1), el proceso de definición de variables en CARTIER-IA depende en gran medida de un flujo de trabajo secuencial y basado en hojas de cálculo.

Los problemas detectados no se limitan a la experiencia de uso, sino que afectan a la arquitectura del sistema, a la capacidad de iterar el modelo de datos, y al propio ciclo de vida del proyecto. A continuación, se analizan en detalle los principales desafíos identificados, comenzando por el enfoque de modelado basado en plantillas.

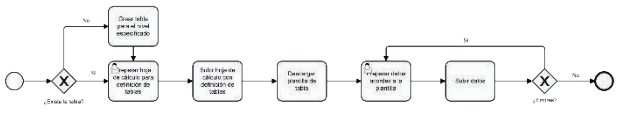


*Figura 1. Flujo de trabajo para la definición de variables y la carga de datos estructurados en CARTIER-IA.*

### 3.1 Modelado basado en plantillas: rigidez y fragilidad

El modelado de variables en CARTIER-IA se articulaba mediante plantillas Excel predefinidas, que describían las entidades y atributos que conformaban el esquema de datos clínicos. Cada hoja de la plantilla correspondía a una tabla de la base de datos (por ejemplo, Paciente, Estudio o Archivo), y cada columna codificaba metadatos con una nomenclatura estricta (SV0, SV1, SV2, … SV6).

- SV0 representaba el nombre de la variable.
- SV1 indicaba el tipo de dato, codificado numéricamente (por ejemplo, 0 = texto, 1 = numérico, 2 = fecha, 3 = booleano).
- SV2–SV6 definían parámetros adicionales, como restricciones, unidades, dominios de valor o reglas de validación.

Este enfoque facilitaba una traducción directa del modelo a la base de datos, reduciendo la necesidad de programación. No obstante, implicaba una dependencia total del formato, de modo que cualquier desviación mínima (como un nombre de hoja incorrecto, un tipo mal codificado o una celda vacía) podía provocar fallos críticos en la carga. Dado que el sistema no incluía validaciones automáticas previas a la importación, los errores solo se detectaban tras ejecutar el proceso de carga, lo que generaba ciclos de corrección largos y frustrantes para los usuarios.

En términos de interacción, este modelo resultaba opaco. Las plantillas actuaban como una "caja negra": el usuario no disponía de retroalimentación inmediata sobre la validez del contenido ni sobre el impacto de sus modificaciones en la estructura global del proyecto. Esto se alejaba de los principios de usabilidad cognitiva aplicados en entornos de modelado visual (como los editores de formularios dinámicos o los lenguajes de dominio específicos), donde las acciones del usuario se acompañan de confirmaciones, ayudas contextuales o vistas previas del resultado.

Otro punto crítico fue la falta de versionado y trazabilidad de los modelos definidos. Cada plantilla representaba un estado estático del esquema, sin un sistema formal para registrar cambios, comparar versiones o revertir configuraciones anteriores. Esto dificultaba el mantenimiento en estudios longitudinales o colaborativos, en los que diferentes investigadores podían modificar de forma independiente las mismas plantillas. En consecuencia, los proyectos tendían a divergir en sus estructuras, comprometiendo la interoperabilidad entre centros.

Además, el enfoque de plantilla limitaba la reutilización de esquemas. Aunque muchos estudios compartían estructuras de datos similares (por ejemplo, información demográfica, medidas clínicas básicas o parámetros de imagen), cada proyecto debía reconstruir sus plantillas desde cero. No existía un repositorio de modelos reutilizables ni un mecanismo para importar subconjuntos de variables validadas previamente. Esto contrastaba con las tendencias actuales en plataformas EDC avanzadas, donde la

modularidad y el versionado semántico de formularios permiten construir rápidamente nuevas bases de datos a partir de componentes existentes.

Desde el punto de vista técnico, la fragilidad del formato también afectaba al flujo de validación y carga. Las rutinas de procesamiento debían comprobar múltiples condiciones para garantizar la coherencia entre hojas, tipos y relaciones, lo que aumentaba la complejidad del código y el riesgo de errores silenciosos. En algunos casos, pequeños cambios en la codificación (por ejemplo, el uso de comas o puntos en campos numéricos, o el formato regional de fechas) provocaban inconsistencias no detectadas hasta la fase de análisis.

Finalmente, el modelo de plantillas condicionaba la curva de aprendizaje de los usuarios. Aunque Excel es familiar para la mayoría de los profesionales clínicos, el uso de convenciones abstractas como "SV0" o "SV5" resultaba poco intuitivo y requería una documentación exhaustiva para evitar malentendidos. En las evaluaciones realizadas (Vázquez-Ingelmo et al., 2021), se observó que persistían dudas sobre el significado de las columnas, las reglas de nomenclatura y las restricciones de valores. Este problema de transparencia semántica se tradujo en una potencial sobrecarga cognitiva significativa, especialmente para usuarios sin experiencia en modelado de datos o bases de datos relacionales.

En resumen, el uso de plantillas estructuradas permitió un avance rápido en la etapa inicial de desarrollo, pero su rigidez y falta de asistencia interactiva limitaron la evolución del sistema. La manipulación directa de las plantillas sin un editor visual guiado ni validaciones inmediatas generó un cuello de botella tanto en la creación de nuevos modelos como en la extensión de los existentes. Esta debilidad constituye uno de los principales impulsores del rediseño propuesto en este trabajo, orientado hacia un modelado estructural asistido y dinámico, con control de versiones, validaciones en tiempo real y soporte para colaboración multiusuario.

### 3.2 Falta de interacción, retroalimentación y visibilidad de las tareas

La evaluación de CARTIER-IA evidenció un desajuste entre la solidez del sistema y el acompañamiento que recibe el usuario durante el modelado, la carga y la validación. Aunque el sistema ejecuta comprobaciones internas de consistencia y tipos, la interfaz ofrece poca retroalimentación inmediata: los procesos se perciben como opacos y, ante un error, el mensaje resultante es a menudo genérico y poco localizado. Esto dificulta identificar qué variable, hoja o regla debe corregirse y alarga las iteraciones.

Asimismo, se observó escasa visibilidad del progreso en operaciones costosas (p. ej., carga masiva de ficheros DICOM), lo que genera incertidumbre sobre si la tarea avanza o ha concluido. La validación en bloque, sin resultados parciales, incrementa el coste de corrección: un detalle menor puede invalidar la carga completa. Finalmente, la ayuda contextual es limitada, por lo que los usuarios deben salir del flujo para consultar documentación externa, con la consiguiente pérdida de continuidad.

Estos problemas no cuestionan la capacidad técnica del sistema, pero sí señalan oportunidades de mejora en la experiencia de uso. En concreto: (i) mensajes explicativos y localizados de validación, (ii) indicadores de progreso y confirmaciones visuales, (iii) validaciones incrementales con resultados parciales y (iv) un historial de actividades que facilite trazabilidad y recuperación. En conjunto, estos elementos permitirían mantener la robustez actual, a la vez que refuerzan la transparencia, el control percibido y la eficacia en equipos multicentro.

### 3.3 Accesibilidad y escalabilidad para usuarios no técnicos

Un aspecto recurrente en el uso de CARTIER-IA fue la brecha entre la potencia técnica del sistema y la accesibilidad percibida por los usuarios no especializados. Aunque la plataforma fue concebida para reducir la dependencia de conocimientos de programación y ofrecer una vía directa para estructurar y validar datos clínicos, en la práctica muchos de sus procesos seguían requiriendo cierto grado de comprensión técnica, especialmente en las fases de configuración inicial y mantenimiento de proyectos.

Las convenciones empleadas en las plantillas (como la codificación de tipos mediante valores numéricos o el uso de etiquetas genéricas) dificultaban la interpretación por parte de perfiles clínicos o de investigación sin experiencia en bases de datos. Este esfuerzo cognitivo añadía una capa de complejidad innecesaria en un entorno donde el foco debía centrarse en la calidad del dato y no en los detalles del formato.

Además, el flujo de trabajo resultaba poco escalable cuando intervenían varios centros o equipos. Cada investigador debía mantener su propia copia de las plantillas, lo que generaba divergencias entre versiones y aumentaba el riesgo de inconsistencias. No existía un control centralizado de esquemas ni mecanismos de bloqueo o fusión de cambios, por lo que la coordinación requería comunicación manual fuera del sistema.

En conjunto, estos factores revelan que, aunque CARTIER-IA demostró ser una solución sólida para integrar datos estructurados e imagen médica, su diseño todavía reflejaba una orientación más técnica que orientada al usuario final. Las limitaciones de accesibilidad, colaboración y escalabilidad no anulaban su potencial, pero sí marcaban el camino hacia un nuevo modelo de interacción más flexible, guiado y distribuido. De ahí surge la necesidad de evolucionar hacia una arquitectura centrada en la experiencia del usuario, con soporte para edición visual y control de versiones, elementos que se abordan en la siguiente sección.

## 4. Propuesta de rediseño y principios del nuevo sistema

La nueva propuesta busca mantener la robustez del *backend* y la compatibilidad con los estándares clínicos, pero incorporando una capa de interacción que facilite la definición, validación y mantenimiento de los modelos de datos sin depender de plantillas rígidas ni conocimientos técnicos avanzados.

El proceso reestructurado se ilustra en la Figura 2. En comparación con la versión anterior (Figura 1), el nuevo flujo de trabajo elimina la necesidad de que los usuarios codifiquen metadatos manualmente o alineen múltiples hojas. En su lugar, la creación de tablas y la definición de variables se integran de forma fluida en la plataforma, seguida de una descarga guiada de la plantilla de datos adecuada.

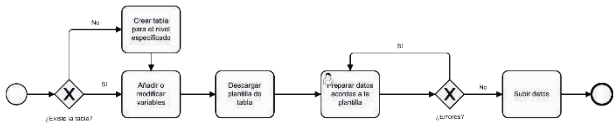



*Figura 2. Flujo de trabajo BPMN actualizado para la carga de datos estructurados mediante el nuevo editor de variables y plantillas validadas.*

### 4.1 Arquitectura interactiva y modelado visual asistido

El nuevo sistema sustituye el flujo basado en plantillas por una arquitectura de modelado visual, en la que el usuario define entidades y atributos directamente desde una interfaz web interactiva. Cada variable se representa como un objeto editable con propiedades asociadas (nombre, tipo, unidades, dominio de valores, restricciones) y con una vista previa inmediata del resultado en el esquema.

Este enfoque reduce la dependencia de formatos externos (como Excel) y evita errores de sintaxis o nomenclatura, ya que las opciones disponibles se adaptan dinámicamente al tipo de dato seleccionado. Además, la interfaz incorpora asistentes contextuales y guías progresivas, que acompañan al usuario en la creación de cada componente, destacando dependencias o incoherencias en tiempo real.

En términos técnicos, el sistema adopta una arquitectura modular basada en microservicios, lo que permite que las operaciones de validación, almacenamiento y exportación se ejecuten de forma independiente y escalable. Esta estructura favorece la extensión futura del sistema, por ejemplo, integrando módulos de inteligencia artificial para sugerir esquemas a partir de conjuntos de datos previos o para detectar redundancias semánticas entre variables.

### 4.2 Validación en tiempo real y retroalimentación

Uno de los pilares del rediseño es la introducción de validaciones dinámicas que operan de forma continua mientras el usuario edita el modelo. En lugar de una validación final y bloqueante, cada campo se verifica al momento, mostrando advertencias localizadas y mensajes interpretables (por ejemplo, “El tipo de dato no coincide con el dominio definido” o “El valor por defecto no pertenece a la lista permitida”).

El sistema incorpora un motor de reglas parametrizable que combina validaciones sintácticas, lógicas y de coherencia intertabla. Esto no solo reduce errores en la carga de datos, sino que también ofrece una retroalimentación pedagógica que guía al usuario en la corrección.

Asimismo, las operaciones largas, como la carga de grandes volúmenes de datos o la sincronización con repositorios DICOM, se acompañan de indicadores de progreso, registros de eventos y confirmaciones visuales, lo que incrementa la transparencia y el control percibido. En caso de error, el sistema proporciona un informe descargable con la descripción y localización exacta de cada incidencia, facilitando la depuración sin romper el flujo de trabajo.

## 5. Implementación y resultados preliminares

El rediseño de la plataforma se ha materializado en una nueva capa de gestión de variables desarrollada con Django, cuyo objetivo es transformar el flujo tradicional de modelado y validación basado en plantillas estáticas en un proceso asistido, guiado e interactivo.

El rediseño de la plataforma no se realizó de manera aislada. Desde las primeras etapas, el desarrollo estuvo guiado tanto por la colaboración continua con el equipo médico que impulsó el proyecto como por la experiencia acumulada durante el uso y la evolución de las versiones previas de CARTIER-IA.

Las reuniones con el equipo clínico permitieron identificar los problemas del flujo basado en hojas de cálculo y las dificultades reales al definir variables o validar datos. A ello se sumó el conocimiento obtenido tras años de uso de CARTIER-IA, donde ya se habían detectado errores recurrentes, puntos de fricción y necesidades no cubiertas.
La combinación de la retroalimentación directa y de la experiencia previa con la plataforma orientó el rediseño hacia un proceso más sencillo, guiado y alineado con la forma de trabajar del personal clínico, asegurando que las decisiones de diseño respondieran a problemas reales.

Esta implementación constituye el núcleo funcional del nuevo sistema: permite definir variables estructuradas, generar automáticamente plantillas Excel con validaciones embebidas y resolver los errores de carga desde la propia aplicación, sin depender de herramientas externas. El enfoque combina tres principios clave:

1. Definición centralizada de variables y restricciones, que garantiza coherencia semántica entre proyectos.
2. Validación temprana en el punto de entrada, a través de reglas automáticas aplicadas en las plantillas generadas.
3. Resolución asistida de errores, que permite corregir incidencias parcial o completamente sin abandonar la interfaz web.

### 5.1 Definición estructurada de variables

El nuevo flujo comienza con la creación de variables desde la interfaz web (Figura 3). Cada variable se define a través de un formulario que incluye campos para su nombre, nivel jerárquico (paciente, estudio o serie), tipo de variable (categórica, binaria, numérica), unidades y restricciones asociadas.

En el caso de variables categóricas, el sistema permite registrar los valores válidos mediante una interfaz interactiva ("Hombre", "Mujer", "N/D", etc.), evitando la necesidad de introducir esta información manualmente en una hoja de cálculo.

De forma análoga, para las variables numéricas, el sistema ofrece la posibilidad de establecer rangos de valores permitidos (por ejemplo, entre 0 y 150 para la edad, o entre 35 y 42 para la temperatura corporal). Estos límites actúan como restricciones dinámicas tanto en la plantilla generada como en la fase de validación de carga: si un valor introducido excede el rango definido, el sistema emite una advertencia inmediata y explica el motivo del rechazo.

Además, el rango puede complementarse con unidades de medida normalizadas (por ejemplo, mmHg, kg, °C), lo que asegura la consistencia semántica y facilita la interoperabilidad entre proyectos. En conjunto, ambos mecanismos trasladan al propio sistema la responsabilidad del control de calidad, reduciendo errores de entrada y homogeneizando la estructura de los datos desde el momento de su creación, convirtiendo las decisiones de modelado en metadatos persistentes, reutilizables en distintos proyectos, y elimina la ambigüedad propia de los esquemas codificados en columnas abstractas.

Las reglas definidas se almacenan en una base de datos relacional y se aplican automáticamente cuando se generan las plantillas de carga o se validan nuevos registros.

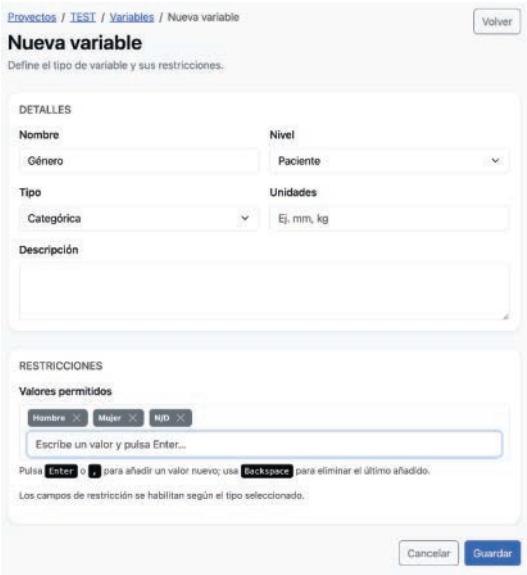



*Figura 3. Interfaz de creación de variables con definición de nivel, tipo y valores permitidos.*

### 5.2 Generación automática de plantillas con validación embebida

Una vez definidas las variables, el sistema genera una plantilla Excel personalizada que incluye las validaciones establecidas en el nivel anterior (Figura 4). Cada columna corresponde a una variable declarada y, para aquellas de tipo categórico o restringido, se crea una lista de selección con los valores permitidos.

Este mecanismo utiliza la validación nativa de Excel para reforzar la calidad del dato desde el punto de entrada: si el usuario intenta introducir un valor fuera de la lista, el sistema muestra un mensaje explicativo, evitando así la inserción de valores inconsistentes.

Además, las plantillas incluyen metadatos ocultos que vinculan cada hoja con su nivel jerárquico (paciente, estudio o serie) y con la versión del modelo de datos, lo que permite detectar discrepancias en la importación y asegurar la trazabilidad de cada variable.


*Figura 4. Validación contextual en Excel generada automáticamente a partir de las restricciones definidas en el sistema.*

### 5.3 Carga y validación

La carga de datos estructurados se realiza íntegramente dentro de la aplicación web (Figura 5). El usuario selecciona el archivo Excel generado y el sistema ejecuta de forma automática una validación estructural y de contenido, que incluye:

- Verificación de identificadores obligatorios y niveles jerárquicos.
- Correspondencia exacta entre las variables declaradas y las columnas presentes.
- Comprobación de tipos, dominios y restricciones de valor.
- Detección de valores ausentes o duplicados.
- Incongruencia con datos existentes.

Los resultados de esta validación se muestran de manera visual e interpretativa. Los errores se agrupan por tipo y nivel (por ejemplo, 4 errores a nivel de Paciente, 0 a nivel de Estudio, 0 a nivel de Serie) y se acompañan de mensajes en lenguaje natural que explican el problema y las posibles acciones.


*Figura 5. Interfaz de carga de datos con validación automática y registro de cargas recientes.*

Una de las innovaciones más relevantes del nuevo flujo es la posibilidad de resolver errores directamente desde la aplicación, sin necesidad de editar la hoja de cálculo manualmente (Figura 5). Tras la validación, el sistema presenta un panel de revisión donde el usuario puede:

- Ignorar o excluir columnas con errores, continuando con la carga parcial de los datos válidos.
- Agregar nuevas variables si se detectan columnas no declaradas (por ejemplo, "Diabetes") y desea incorporarlas al modelo.
- Reintentar la carga tras corregir la plantilla.

De esta forma, el flujo de trabajo se mantiene dentro del entorno, reduciendo drásticamente las interrupciones y la dependencia de soporte técnico.

### 5.4 Historial de validaciones y trazabilidad del proceso

Todas las operaciones de carga quedan registradas en un historial de validaciones (Figura 7), donde se detallan la fecha, el usuario, el estado final y el número de errores detectados. Cada entrada permite acceder al informe completo sin necesidad de volver a subir la plantilla.

Este registro garantiza la trazabilidad completa del proceso y facilita la auditoría de las acciones realizadas sobre cada conjunto de datos.


*Figura 7. Historial de validaciones con registro de estado, usuario y número de errores.*

## 6. Discusión

El rediseño propuesto aborda directamente varios de los problemas de usabilidad y flujo de trabajo identificados durante la evaluación heurística de la versión original de CARTIER-IA (Vázquez-Ingelmo et al., 2021). Al integrar la definición de variables en una interfaz web guiada y automatizar la generación de plantillas validadas, el sistema traslada la complejidad desde el usuario final hacia la lógica interna de la plataforma, donde puede ser controlada, auditada y evolucionada con mayor seguridad.

El proceso de rediseño también se vio reforzado por la experiencia de uso de versiones previas de CARTIER-IA y por el intercambio continuo con el equipo médico involucrado en el proyecto, lo que permitió ajustar la solución a problemas y patrones de trabajo observados en la práctica.

Este enfoque mejora la conformidad con múltiples heurísticas de Nielsen (Nielsen, 1992), especialmente las siguientes:

- **HR1 (Visibilidad del estado del sistema):** los usuarios reciben retroalimentación visual inmediata durante la creación de variables y pueden descargar plantillas con la garantía de que coinciden con el esquema interno del proyecto.
- **HR5 (Prevención de errores):** las validaciones se aplican tanto en la interfaz como en los archivos Excel generados, reduciendo la posibilidad de errores de introducción o de interpretación.
- **HR6 (Reconocimiento mejor que recuerdo):** el diseño minimiza la necesidad de memorizar códigos o convenciones estructurales, mostrando las opciones de configuración de forma contextual.
- **HR9 (Ayuda al reconocimiento, diagnóstico y recuperación de errores):** los errores de plantilla se previenen por construcción, y las futuras versiones podrán incorporar retroalimentación integrada durante la carga de datos.
- **HR10 (Ayuda y documentación):** los elementos visuales y la propia estructura de la interfaz reducen la necesidad de consultar documentación externa, simplificando la incorporación de nuevos usuarios.

Más allá de estas heurísticas, los beneficios se extienden a la experiencia general de uso. Al alinear el modelado de variables con flujos de trabajo intuitivos, el rediseño reduce el tiempo de aprendizaje, disminuye la carga de soporte técnico y facilita la colaboración entre perfiles técnicos y clínicos. Además, la separación clara de responsabilidades, donde los expertos de dominio definen las variables y el sistema garantiza la estructura, representa un modelo más sostenible y escalable para la gestión de datos estructurados en investigación médica.

En términos de extensibilidad, la nueva arquitectura abre oportunidades adicionales. Por ejemplo, las variables podrían anotarse con etiquetas semánticas como *SNOMED* (Spackman et al., 1997) o *LOINC* (McDonald et al., 2003) para favorecer la interoperabilidad con otros sistemas. También podría integrarse un mecanismo de versionado que registre los cambios en las definiciones a lo largo del tiempo, o un repositorio central de conjuntos reutilizables de variables que facilite la armonización entre proyectos.

De manera complementaria, el cambio hacia un modelado guiado desde la plataforma sienta las bases para incorporar funcionalidades avanzadas como la creación de variables asistida por inteligencia artificial. Analizando esquemas previamente definidos o patrones de datos, el sistema podría recomendar variables, detectar redundancias o sugerir estrategias de codificación óptimas para los análisis posteriores.

Otro de los beneficios destacados del nuevo sistema es la posibilidad de reutilizar definiciones de variables entre proyectos. Las variables definidas en un contexto pueden clonarse o importarse fácilmente a otro, reduciendo la redundancia y acelerando la configuración de nuevos estudios. Esto resulta especialmente útil en entornos multicéntricos o longitudinales, donde la coherencia de las definiciones es esencial.

Finalmente, las variables pueden agruparse en categorías o dominios semánticos (por ejemplo, demografía, comorbilidades, características de imagen), lo que ayuda a organizar esquemas extensos de forma más efectiva y comprensible.

No obstante, persisten algunas limitaciones. Aunque el rediseño simplifica la definición y validación, el sistema sigue dependiendo de la introducción de datos mediante hojas de cálculo, lo que conlleva ciertos riesgos: el usuario aún puede modificar accidentalmente encabezados, alterar codificaciones o perder reglas de validación si utiliza herramientas no compatibles. Además, la plantilla continúa desconectada de la lógica interna durante la fase de entrada de datos, sin ofrecer orientación en tiempo real ni edición colaborativa.

En cuanto a la evaluación, la ausencia de estudios formales con usuarios en esta fase responde a una decisión deliberada. Antes de realizar una validación empírica, era necesario estabilizar la nueva arquitectura y comprobar que los cambios propuestos resolvían los principales puntos de fricción detectados en el sistema previo. El trabajo actual proporciona esa base técnica y conceptual, preparando el camino para estudios de usabilidad más completos en etapas posteriores.

Aun así, el trabajo representa una transición clara de una visión centrada en la herramienta a una centrada en el usuario, redefiniendo el modelado de datos estructurados como un proceso cooperativo, comprensible e iterativo. Esta

evolución sienta un precedente para futuras versiones de este sistema y para otras plataformas de investigación clínica que busquen infraestructuras más accesibles, robustas e inteligentes.

## 7. Conclusiones y trabajos futuros

Este trabajo presenta un rediseño centrado en el usuario del flujo de definición de variables en entornos clínicos, una plataforma que integra la gestión de datos estructurados y de imágenes médicas para la investigación clínica.

La nueva solución sustituye el proceso tradicional propenso a errores y basado en hojas de cálculo por una interfaz web interactiva que permite definir, categorizar y reutilizar variables de manera más sencilla y fiable.

En paralelo, el sistema genera automáticamente plantillas de entrada de datos validadas, reduciendo errores de formato y asegurando el cumplimiento de las restricciones semánticas definidas en el modelo.

El flujo de trabajo actualizado y la nueva interfaz abordan directamente los principales problemas de usabilidad identificados en evaluaciones previas, especialmente aquellos relacionados con la prevención de errores, la retroalimentación al usuario y la documentación contextual.

Al incorporar conocimiento de dominio y reglas de validación tanto en la interfaz como en las plantillas generadas, el sistema disminuye la carga cognitiva y mejora la accesibilidad para profesionales clínicos y colaboradores sin perfil técnico.

De cara al futuro, el trabajo seguirá dos líneas principales. Primero, la realización de pruebas de usabilidad, mediante nuevos estudios que combinen análisis heurísticos y evaluaciones centradas en el usuario para medir cuantitativamente la eficiencia, precisión y satisfacción comparadas con el flujo original. Segundo, integración de herramientas de inteligencia artificial asistida, capaces de sugerir nombres, tipos y restricciones de variables a partir de esquemas previos o descripciones textuales de estudios. Estas funcionalidades buscan agilizar la configuración inicial de proyectos y promover la reutilización de patrones validados entre investigaciones.

## Agradecimientos

Esta investigación fue financiada parcialmente por el Ministerio de Ciencia e Innovación de España a través del proyecto AVisSA, número de referencia PID2020-118345RB-I00. Este trabajo también fue apoyado por ayudas competitivas comunitarias (GRS 2033/A/19, GRS 2030/A/19, GRS 2031/A/19, GRS 2032/A/19) del SACYL, Junta de Castilla y León; por ayudas competitivas nacionales (PI14/00695, PIE14/00066, PI17/00145, DTS19/00098, PI19/00658, PI19/00656, PI21/00369) del Instituto de Salud Carlos III, Ministerio de Ciencia e Innovación de España, cofinanciadas por FEDER/FSE, "Una manera de hacer Europa"; y por el CIBERCV (CB16/11/00374) del Instituto de Salud Carlos III, Ministerio de Ciencia e Innovación de España.

## Declaración sobre el uso de IA generativa

Durante la preparación de este trabajo, los autores utilizaron ChatGPT y Grammarly para: revisión gramatical y ortográfica, parafraseo y reformulación de textos. Tras utilizar estas herramientas/servicios, los autores revisaron y editaron el contenido según fue necesario y asumen la plena responsabilidad del contenido de la publicación.

## Referencias

Dobell, E., Herold, S., & Buckley, J. (2018). Spreadsheet error types and their prevalence in a healthcare context. *Journal of Organizational and End User Computing, 30*, 20–42.

García-Peñalvo, F. J., Vázquez-Ingelmo, A., García-Holgado, A., Sampedro-Gómez, J., Sánchez-Puente, A., Vicente-Palacios, V., Dorado-Díaz, P. I., & Sánchez-Fernández, P. L. (2021). Application of artificial intelligence algorithms within the medical context for non-specialized users: The CARTIER-IA platform. *International Journal of Interactive Multimedia and Artificial Intelligence, 6*, 46–53.

García-Peñalvo, F. J., Vázquez-Ingelmo, A., García-Holgado, A., Sampedro-Gómez, J., Sánchez-Puente, A., Vicente-Palacios, V., Dorado-Díaz, P. I., & Sánchez, P. L. (2024). Koopaml: A graphical platform for building machine learning pipelines adapted to health professionals. *International Journal of Interactive Multimedia and Artificial Intelligence*.

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics, 42*, 377–381.

Herrick, R., Horton, W., Olsen, T., McKay, M., Archie, K. A., & Marcus, D. S. (2016). XNAT Central: Open sourcing imaging research data. *NeuroImage, 124*, 1093–1096.

Iyengar, S. P., Acharya, H., & Kadam, M. (2019). Big data analytics in healthcare using spreadsheets. En *Big Data Analytics in Healthcare* (pp. 155–187). Springer.

McDonald, C. J., Huff, S. M., Suico, J. G., Hill, G., Leavelle, D., Aller, R., Forrey, A., Mercer, K., DeMoor, G., Hook, J., et al. (2003). LOINC, a universal standard for identifying laboratory observations: A 5-year update. *Clinical Chemistry, 49*, 624–633.

Nielsen, J. (1992). Finding usability problems through heuristic evaluation. En *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 373–380).

Spackman, K. A., Campbell, K. E., & Côté, R. A. (1997). SNOMED RT: A reference terminology for health care. En *Proceedings of the AMIA Annual Fall Symposium* (p. 640).

Vázquez-Ingelmo, A., Alonso, J., García-Holgado, A., García-Peñalvo, F. J., Sampedro-Gómez, J., Sánchez-Puente, A., Vicente-Palacios, V., Dorado-Díaz, P. I., & Sánchez, P. L. (2021). Usability study of CARTIER-IA: A platform for medical data and imaging management. En P. Zaphiris & A. Ioannou (Eds.), *Learning and Collaboration Technologies: New Challenges and Learning Experiences* (pp. 374–384). Springer.

INTERACCION
Revista Digital de AIPO
Asociación Interacción Persona-Ordenador
Vol. 6, No 2 (2025)
ISSN 2695-6578

# Interacción natural y emocional con robots sociales

## Natural and emotional interaction with social robots

**Liany Mendoza**
Departamento de Informática e Ingeniería de Sistemas, Universidad de Zaragoza, Zaragoza, España
lmendoza@unizar.es

**Eva Cerezo**
Departamento de Informática e Ingeniería de Sistemas, Universidad de Zaragoza, Zaragoza, España
ecerezo@unizar.es

**Loreto Matinero**
Departamento de Informática e Ingeniería de Sistemas, Universidad de Zaragoza, Zaragoza, España
lmatinero@unizar.es

**Adrián Arribas**
Departamento de Informática e Ingeniería de Sistemas, Universidad de Zaragoza, Zaragoza, España
795593@unizar.es

**Sandra Baldassarri**
Departamento de Informática e Ingeniería de Sistemas, Universidad de Zaragoza, Zaragoza, España
sandra@unizar.es

Recibido: 15.11.2025 | Aceptado: 01.12.2025

**Palabras Clave**

Robots sociales
Computación afectiva
LLM
Interacción natural
Interacción emocional

**Resumen**

En los últimos años, se ha demostrado la gran cantidad de beneficios que los robots sociales pueden ofrecer a las personas, tanto como herramientas de aprendizaje como en funciones de acompañamiento social. Para que estas interacciones sean efectivas, es necesario que los robots sean capaces de conversar de forma inteligente y actúen de forma natural y emocionalmente coherente. En este trabajo se presenta un sistema orientado a mejorar la interacción humano-robot mediante la integración de herramientas de inteligencia artificial basada en modelos de lenguaje (LLM), interfaces gráficas (visual y táctil), movimientos, luces y voz. El sistema, implementado en Android sobre el robot Sanbot Elf, se compone de tres módulos principales: el Módulo Conversacional, el Módulo Emocional y el Módulo Reactivo, diseñados para lograr un diálogo fluido, expresar estados emocionales y aportar naturalidad al comportamiento del robot. Se realizó una evaluación inicial del sistema conversacional con 18 usuarios, cuyos resultados fueron muy positivos, y dieron lugar a una serie de mejoras para optimizar el desempeño y la experiencia de interacción.

**Keywords**

Social robots
Affective computing
LLM
Natural interaction
Emotional interaction

**Abstract**

In recent years, social robots have been shown to offer a wide range of benefits to people, both as learning tools and in social companionship roles. For these interactions to be effective, robots need to be able to converse intelligently and act in a natural and emotionally coherent manner. This paper presents a system aimed at improving human-robot interaction through the integration of artificial intelligence tools based on language models (LLM), graphical interfaces (visual and tactile), movements, lights, and voice. The system, implemented in Android on the Sanbot Elf robot, consists of three main modules: the Conversational Module, the Emotional Module, and the Reactive Module, designed to achieve fluid dialogue, express emotional states, and bring naturalness to the robot's behavior. An initial evaluation of the conversational system was carried out with 18 users, the results of which were very positive and led to a series of improvements to optimize performance and interaction experience.

## 1. Introducción

Los robots sociales están adquiriendo una presencia cada vez más significativa en la sociedad actual. Más allá del uso de robots para trabajo y asistencia en tareas industriales que requieren alta precisión y exactitud, los robots sociales están comenzando a integrarse en la vida cotidiana como asistentes, ayudantes, tutores e incluso compañeros (Van y Reed, 2010). Aunque las definiciones de los robots sociales son heterogéneas (Henschel, Laban y Cross, 2021), en el trabajo de Sarrica, Brondi y Fortunati (2020) se identificaron ciertos rasgos comunes: los robots sociales son agentes corpóreos que poseen distintos grados de autonomía y participan en interacciones sociales con humanos,

comunicándose, cooperando y tomando decisiones. Este componente social es fundamental para que los robots proporcionen apoyo físico y emocional, manteniendo interacciones con las personas a largo plazo (Churamani, Kalkan y Gunes, 2020).

A pesar de que los robots no igualan las habilidades humanas en la interacción social (Cross, Hortensius y Wykowska, 2019), diversos estudios destacan su potencial cuando se introducen de manera contextualizada y ética (Wullenkord y Eyssel, 2020), mejorando la calidad de vida y promoviendo el bienestar (Yang et al., 2020). Una base de pruebas cada vez mayor documenta cómo los robots sociales podrían funcionar como herramientas autónomas de apoyo en intervenciones cognitivas (Alnajjar et al., 2019), rehabilitación física y fisioterapia (Assad-Uz-Zaman et al., 2019; Chen, García-Vergara y Howard, 2018; Mohebbi, 2020), intervenciones psicosociales y mejoras en el bienestar (Robinson, Cottier y Kavanagh, 2019; Scoglio et al., 2019), así como en el ámbito educativo y del entretenimiento (Kyprianou et al., 2023; Calvo-Barajas, Perugia y Castellano, 2020). En los últimos años, los robots sociales diseñados para acompañar a las personas se han vuelto más comunes, especialmente en contextos de atención a mayores o niños siendo de gran importancia para su aceptación, que estos robots sociales sean capaces de mantener conversaciones naturales y empáticas, y tengan en cuenta aspectos emocionales. Este tipo de sistemas ha crecido gracias a la integración de modelos de IA y del procesamiento del lenguaje natural (Spezialetti, Placidi y Rossi, 2020). La naturalidad conversacional es un elemento esencial para lograr interacciones significativas entre humanos y robots (Romat et al., 2016). Las técnicas modernas de IA, especialmente los modelos de lenguaje de gran tamaño (LLM), permiten mantener conversaciones más fluidas y contextualmente coherentes, considerando emociones, tonos y contextos del interlocutor (Laban y Cross, 2024). Asimismo, la expresividad corporal y la comunicación no verbal son fundamentales para la credibilidad y eficacia de la interacción (Breazeal, 2003). En este trabajo se propone el desarrollo de un sistema conversacional basado en un modelo de lenguaje de gran tamaño (LLM), que aprovecha las capacidades de interacción multimodal de un robot social (Sanbot Elf) que soporta la interacción natural, emocional y personalizada humano-robot. Además, se ha considerado importante que el robot actúe de forma natural imitando el comportamiento humano mediante gestos y comportamientos no verbales.

A continuación, en el apartado 2, se realiza el estado del arte donde se abordan las principales características que un robot debe tener para lograr una buena interacción entre el humano y el robot. En el apartado 3 se describe el robot, las consideraciones de diseño e implementación del sistema conversacional; mientras que el apartado 4 se ofrece un ejemplo de conversación entre una persona y el robot. Luego, en el apartado 5, se describe la evaluación llevada a cabo con usuarios y las mejoras realizadas en base a los resultados obtenidos. Por último, se finaliza con las conclusiones de la investigación y el trabajo futuro.

## 2. Estado del arte

Para lograr una conversación natural y emocional entre humanos y robots es necesario comprender cómo un robot puede desarrollar comportamientos sociables. Para ello, se ha realizado una revisión de la literatura, con el propósito de identificar y analizar las características de comunicación y sociabilidad que deben integrarse en el diseño y funcionamiento de los robots sociales que permitan una interacción emocional y natural cuando conversan con las personas. Los modelos lingüísticos grandes (LLM) son excelentes para mantener una conversación fluida debido a su capacidad para comprender el contexto e interpretar comandos de lenguaje natural (Wang et al., 2024). En este sentido, se están introduciendo como una herramienta fundamental de soporte a la conversación humano-robot (Kim, Lee y Mutlu, 2024; Cherakara et al., 2023) debido a su gran potencial en el desarrollo de aplicaciones conversacionales capaces de expresar y reconocer emociones humanas, ampliando las fronteras de la interacción humano-robot (Jiang et al., 2025). En cuanto a su uso en robots sociales, Rawal et al. (2024) proponen un agente multimodal que integra información visual y textual para generar respuestas empáticas, combinando las expresiones faciales del usuario utilizando el modelo LLaMA2 para que el sistema reconozca y refleje estados emocionales, generando interacciones naturales y afectivas. Recientemente, Pinto-Bernal, Biondina y Belpaeme (2025) demuestran que la incorporación de LLMs en robots sociales posibilita interacciones más fluidas, coherentes y culturalmente adaptadas, al incluir mecanismos de memoria conversacional, ajuste de estilo y adaptación multilingüe. Ambos trabajos evidencian que estos modelos no solo mejoran la coherencia lingüística y contextual de los diálogos, sino que también permiten diseñar sistemas conversacionales emocionalmente inteligentes, capaces de establecer vínculos sociales más cercanos y significativos con los usuarios.

Sin embargo, no es suficiente con que el robot sea capaz de sostener una conversación verbal, sino que también debe emplear mecanismos de comunicación no verbal, tales como las emociones y las expresiones faciales (Bonarini, 2020). En

diferentes contextos, como el educativo se ha observado (Kyprianou et al., 2023) que los estudiantes esperan que los robots puedan entender y expresar emociones como alegría, tristeza, miedo y vergüenza. Esto incluye la capacidad de mostrar emociones a través de expresiones faciales, tono de voz y movimientos de manos, y que puedan interactuar de manera similar a un amigo humano, incluyendo la capacidad de hablar y reaccionar al tacto. Así, en la interacción entre robots y niños se ha visto (Calvo-Barajas, Perugia y Castellano, 2020) que las expresiones faciales de felicidad y enojo de los robots influyen en la percepción de confianza, simpatía y competencia por parte de los niños durante los primeros encuentros. Pero esto no se da solo en contextos educativos, según un estudio de De Graaf, Allouch y Van Dijk (2015), la capacidad de un robot para mantener un diálogo, analizar el estado de ánimo de su usuario y expresar emociones son las características claves necesarias para que un robot social sea aceptado. También es muy importante que la emoción expresada por un robot esté más orientada hacia los demás que hacía sí mismo: la empatía (Paiva, 2017) constituye un aspecto clave en la interacción entre humanos y robots (Brave, Nass y Hutchinson, 2005; Leite et al., 2013).

Para lograr estos objetivos, es fundamental el estudio e incorporación de componentes emocionales (Gou et al., 2014), ya que los robots que manifiestan emociones mediante gestos o expresiones faciales tienden a ser percibidos como más cercanos y familiares. Además, la capacidad expresiva del robot facilita la comprensión de su comportamiento, estado emocional, motivación y razonamiento (Xu et al., 2015), lo que, a su vez, contribuye a que sea visto como confiable y empático. Diversos estudios demuestran, además, que se espera que los robots sociales sean sensibles a las diferencias individuales (debidas a la cultura, la edad, el sexo o la personalidad, entre otros factores), ofreciendo una experiencia de interacción natural y atractiva personalizada para cada usuario (Lee et al, 2012; Dziergwa et al., 2018). Por tanto, aunque es esencial mejorar la funcionalidad de los robots sociales, también es importante que el robot sea capaz de comunicarse de una forma que se considere socialmente aceptable (De Graaf, Allouch y Van Dijk, 2015). Por otra parte, es importante también que el robot tenga un comportamiento similar al de un humano (Melo y Moreno, 2022), por lo que debe ser capaz de reaccionar en tiempo real a estímulos del entorno durante una conversación (Reimann et al., 2024). Esta combinación de elementos verbales y no verbales (Babel et al., 2021) contribuye a la fluidez y naturalidad en la comunicación entre humanos. Por esta razón, en el diseño de la interacción humano-robot resulta esencial la implementación de técnicas multimodales, señales verbales y no verbales, y comportamientos naturales o reactivos para lograr una comunicación real entre humanos y robots.

Los estudios mencionados previamente abordan, de forma separada, diferentes aspectos. Por ejemplo, algunos trabajos se centran en la comunicación y el potencial de los modelos de lenguaje de gran escala (LLM) (Kim, Lee y Mutlu, 2024), incluso incluyendo expresiones faciales (Cherakara et al., 2023; Rawal et al., 2024). Otros trabajos se focalizan en estudiar la interacción, la aceptación y la confianza que generan los robots cuando son capaces de expresar emociones (Bonarini, 2020; Calvo-Barajas, Perugia y Castellano, 2020; Kyprianou et al., 2023). Y, en otros trabajos, se centran en que el robot presente comportamientos y reacciones similares a las humanas (Babel et al., 2021; Melo y Moreno, 2022; Reimann et al., 2024). Sin embargo, estos estudios, a diferencia del que se presenta en este trabajo, no integran en un único sistema la conversación fluida basada en IA, la expresión emocional a través del movimiento, el cuerpo y las expresiones faciales, la detección y reconocimiento de estímulos externos y las reacciones y comportamientos físicos naturales.

## 3. Descripción del robot y consideraciones de diseño del sistema conversacional

En este apartado se analizan los componentes esenciales que permiten desarrollar una interacción conversacional, emocional y natural entre el ser humano y un robot social. Para ello, se presenta en primer lugar el robot Sanbot Elf y sus características esenciales, como plataforma de interacción. Dado el importante papel que desempeña la inteligencia artificial generativa, y en particular los modelos de lenguaje de gran escala (LLMs), se realiza un análisis comparativo de distintos modelos LLM con el fin decidir el modelo más óptimo para nuestro desarrollo. En tercer lugar, se aborda la definición y generación de reacciones naturales en Sanbot Elf como elemento clave para una interacción más natural y amigable.

### 3.1 El robot Sanbot Elf

El robot Sanbot Elf, ofrece una serie de funcionalidades claves para el desarrollo del módulo conversacional, aunque también presenta algunas limitaciones. Entre sus características destacadas se encuentra el control del habla, que incluye funciones de Text to Speech (TTS) para convertir texto en audio con ajustes de velocidad y entonación, y Speech to Text (STT) para reconocer y transformar el habla del usuario en texto, facilitando la interpretación del lenguaje natural. En cuanto al control del movimiento, Sanbot Elf (mide 90,2 cm) puede mover su cabeza, brazos y

ruedas omnidireccionales, características fundamentales para expresar emociones mediante gestos y movimientos, así como controlar LEDs de color en la cabeza, brazos y base (ver Figura 1). Por otro lado, el control del audio permite ajustar el volumen de los altavoces y gestionar el micrófono, optimizando la interacción verbal. Además de la interacción verbal, Sanbot permite la interacción táctil a través de una pantalla de 10.1 pulgadas que posee en su frontal y dispone de un gran número de sensores de tacto en distintas partes de su cuerpo (ver Figura 1), además de dos cámaras que abren la posibilidad a reconocer expresiones faciales y de rostros. Las mayores limitaciones vienen de su movilidad, ya que no dispone de módulos de navegación. El robot también dispone de una pequeña pantalla, a modo de cara, que permite mostrar 18 expresiones faciales diferentes (ver Figura 2). En cuanto al software, los fabricantes implementaron un SDK, llamado QihanOpenSDK, el cual fue utilizado para el desarrollo de la herramienta conversacional que se presenta en la arquitectura.

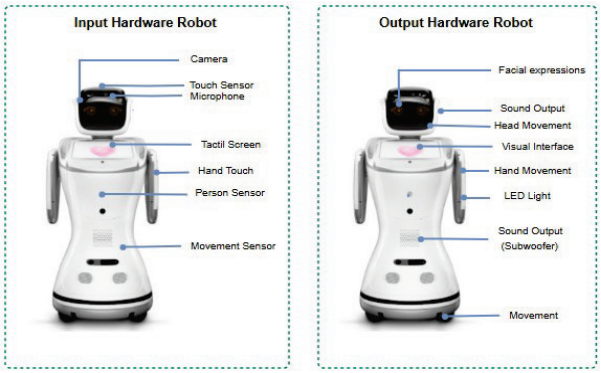



*Figura 1: Componentes físicos del robot Sanbot Elf.*


*Figura 2: Expresiones faciales del robot Sanbot Elf*

### 3.2 Determinación de la herramienta de IA para conversaciones fluidas

Para lograr una conversación fluida y natural entre el robot y las personas es necesario generar respuestas coherentes y contextualizadas, y luego reproducir las respuestas mediante voces artificiales más cercanas a la voz humana. Con este propósito, se llevó a cabo una comparativa de diferentes modelos disponibles en mercado, evaluando sus capacidades: procesamiento lingüístico y síntesis de voz natural. En primer lugar, se decidieron descartar herramientas cómo: Speechify, Murf AI, Wondercraft, ElevenLabs o Lovo, las cuales sí permitían la reproducción de texto usando TTS (Text to Speech), pero no eran LLMs (Large Language Model). En la Tabla 1 se presentan otras herramientas LLM analizadas.

*Tabla 1: Comparación de herramientas de Inteligencia Artificial*

| Herramienta | TTS | Observaciones |
|---|---|---|
| LLaMA | Sí | Requiere un buen equipo o servidor adecuado. |
| Google Cloud | Sí | Plataforma con muchas API confusas. |
| OpenAI | Sí | Precios variables según el modelo LLM utilizado. |

Para la selección definitiva se descartó LLaMA, por no contar con los recursos necesarios para su instalación y Google Cloud debido a su confusa biblioteca con numerosas API, quedando la herramienta de OpenAI como la más óptima para esta investigación. Esta herramienta ofrece una muy alta calidad y consistencia en cuanto a sus modelos de diálogo y razonamiento, además de ser fácil de usar, estar bien documentada y no necesitar infraestructura propia. Una vez seleccionada **OpenAI**, fue necesario determinar qué modelo de LLM resultaba más adecuado para los objetivos del trabajo. Para ello, se realizó una comparativa de las ventajas y desventajas de los diferentes modelos disponibles, entre ellos GPT-3.5, GPT-4, y GPT-4o mini, considerando factores como precisión, velocidad de respuesta, capacidad de razonamiento, consumo de recursos y costo. Finalmente, se decidió seleccionar **GPT-4o mini**, un modelo optimizado que mantiene el rendimiento y la calidad de razonamiento del modelo GPT-4o (versión multimodal de GPT-4), pero con un menor consumo computacional y una velocidad de inferencia significativamente mayor. Este modelo tiene un menor coste y un buen rendimiento, lo que lo hace especialmente apropiado para entornos donde se requiere procesamiento rápido y respuestas coherentes. Además, GPT-4o mini es compatible con tareas multimodales (texto, imagen, audio y video) y su entrenamiento se ha realizado con datos multilingües lo que permite realizar con calidad, diálogos prolongados. Además, OpenAI cuenta con funciones Text to Speech, con la posibilidad de realizar la petición enviando un prompt personalizado, ofreciendo al cliente dos posibles modelos: TTS y TTS-HD. Ya que se buscaba la menor latencia posible y la diferencia de calidad es apenas apreciable, se seleccionó el modelo TTS estándar.

### 3.3 Diseño de reacciones naturales en Sanbot Elf

Para mejorar el dinamismo de la conversación y humanizar la actuación del robot, es importante incluir movimientos propios del comportamiento humano, como el lenguaje corporal o las reacciones espontáneas durante la conversación entre la persona y el robot. En este caso, y teniendo en cuenta las posibilidades de movilidad y reconocimiento de las que dispone el robot Sanbot Elf, se ha considera importante diseñar e incorporar las siguientes funcionalidades:

**Ejecución de movimientos naturales:** Esta funcionalidad permite incluir una serie de movimientos los cuales no dependan de ningún estímulo externo, y permitan al robot moverse de forma natural y aleatoria durante una conversación.

**Detección de ruido:** Esta funcionalidad permite al robot reaccionar reactivamente ante ruidos que se producen en el entorno, mediante una queja o pidiendo silencio.

**Localización de sonido:** Esta función permite al robot reaccionar frente a una fuente de sonido moviendo su cabeza hacia la posición del sonido percibido.

**Detección de personas:** Esta funcionalidad es la que va a permitir al robot reaccionar en el caso de que una persona pase por delante o detrás suyo. La reacción consiste en una frase aleatoria como: "¡Hola! ¿Te apetece hablar?" o "¡Hola! ¿Qué tal estas?", o en el caso de que la persona pase por detrás el robot girará su cuerpo 180º y reproducirá una frase de las siguientes: "¡Me asustaste! No te había visto", "¡Ah! Debería tener ojos en la espalda" o "Hola, parece que me estabas espiando".

**Reconocimiento facial:** Esta función permite al robot realizar detecciones faciales (localización de una cara en la imagen capturada), reconocimiento de expresiones faciales, reconocimiento de edad y género, y reconocimiento facial (reconocimiento de una persona concreta), y reaccionar ante esa detección. Por ejemplo, si el robot detecta una expresión de enfado reproducirá una frase como frase: "¡Eh! No te enfades conmigo, yo no tengo la culpa".

### 3.4 Implementación del sistema

En la Figura 3 se muestra el sistema desarrollado, que consta de tres bloques.
El primer bloque corresponde a las Interfaces de Usuario, donde se incluyen pantallas para la interacción con el robot, como la pantalla conversacional, la de configuración y personalización. El segundo bloque es el Módulo Conversacional, que constituye el núcleo de la interacción y está compuesto por dos elementos clave: el Módulo Emocional, encargado de gestionar las emociones del robot, asegurando una respuesta más natural y expresiva; y el Módulo de Diálogo, responsable de la gestión de la conversación, construyendo los mensajes intercambiados entre el usuario y el robot. El tercer bloque engloba componentes externos utilizados en la implementación, como el Módulo de IA, encargado del procesamiento del lenguaje natural mediante reconocimiento de voz y generación de respuestas, las APIs Externas, que permiten realizar consultas y acceder a funciones adicionales; y el Control del Robot, que gestiona hardware como el movimiento de extremidades, el habla y el sistema en general. A continuación, se comentan los componentes principales: Interfaces de usuario, Módulo conversacional (que incluye gestor de dialogo, modulo emocional y modulo reactivo), y finalmente se presenta la Integración con otros

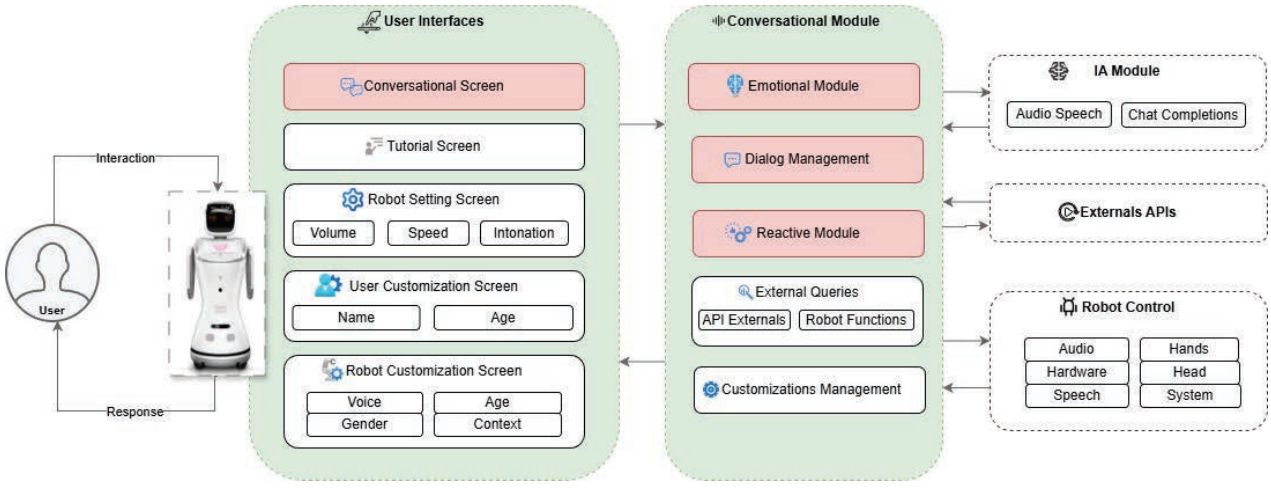



*Figura 3: Sistema conversacional desarrollado (sombreados en color rojo los componentes claves de la herramienta).*

componentes (módulo de IA y control del robot).

### *3.4.1 Interfaces de Usuario*

Consta de tres pantallas principales (ver Figura 4):

- Pantalla Conversacional (*Conversational Screen*): Es la interfaz donde el usuario interactúa y visualiza la conversación mantenida con el robot. El usuario puede iniciar conversaciones, repetir lo que el robot ha dicho o detener su habla.
- Pantalla Personalización del Usuario (*User Customization Screen*): El usuario puede introducir información personal como nombre y edad, datos que se utilizaran durante la conversación para personalizar la interacción, mencionar que esta pantalla es opcional ya que el robot comienza la conversación preguntando dichos datos en caso de que no hayan sido introducidos anteriormente.
- Pantalla Personalización del Robot (*Robot Customization Screen*): Permite al usuario ajustar características del robot: el tipo de voz, género simulado y grupo de edad, lo que influye en cómo el robot interactúa con el usuario.

Además, existen otras dos pantallas, del Tutorial (*Tutorial Screen*) y la de ajuste generales del Robot (*Robot Setting Screen*) que permite al usuario modificar parámetros como el volumen, la velocidad y la entonación de la voz del robot.

### *3.4.2 Módulo Conversacional*

Destacan tres componentes:

- Gestor de Diálogo (*Dialogue Management*): Gestiona la interacción del chat entre el usuario y el modelo LLM (*Large Language Model*). Este componente es el encargado de crear las burbujas de diálogo para que la interfaz visual de conversación pueda mostrar el hilo de la conversación.
- Módulo Emocional (*Emotional Module*): Módulo que gestiona las emociones del robot y las interacciones emocionales con el usuario. A partir de las emociones detectadas por la LLM y del reconocimiento emocional facial, dicho módulo gestiona la salida emocional del robot. Para ello utiliza controles del robot (manos, cabeza, hardware y sistema) para expresar emociones a través de movimientos, luces LED y expresiones faciales (ver Figura 2). El módulo emocional se encarga de hacer el mapeo entre las emociones detectadas en el discurso del usuario y las posibles salidas del robot. Se utilizó las seis expresiones de Ekman (2014), lo que implicó un mapeo al catálogo de expresiones faciales disponibles en el robot. En relación con los colores LED del robot, se tomó como referencia el código de colores mencionado por E. Son (2022), referenciando la película de Disney *Inside Out*. Se consideró que este esquema cromático sería fácilmente reconocible, facilitando así la interpretación emocional del robot.
- Módulo Reactivo (Reactive Module): Módulo que recibe como entrada diversos estímulos del entorno que le rodea, como la percepción y movimientos de los usuarios que están conversando con el robot, o la voz o sonidos capturados del entorno. Las respuestas a estos estímulos generan las salidas de este módulo reactivo, que consisten principalmente en que el robot ejecute acciones de movimiento o de habla acordes con la situación percibida. Para esto, se han incorporado las cinco funcionalidades mencionadas en el apartado anterior (ver Figura 5), y las cuales funcionan de la siguiente forma:

**Reconocimiento facial:** Se ha utilizado el módulo de gestor de multimedia, el cual procesa la entrada


*Figura 4: Pantallas: (1) Personalización del usuario (2) Conversación (3) Personalización del robot*

de vídeo capturada por la cámara HD del robot y el módulo de visión desarrollado para agentes sociales interactivos (Ang et al., 2024). Para estas funciones el módulo reactivo recibe del módulo de visión la información concreta sobre las coordenadas del cuadro delimitador del rostro (detección facial), una de las seis expresiones de Ekman (2014) (reconocimiento de expresiones faciales), la edad y el género (reconocimiento de edad y género), y el nombre de la persona reconocida (reconocimiento de una persona) según un conjunto predefinido de usuarios. En todos los casos el robot reacciona mediante la función de habla, incluyendo en sus frases la información detectada.

**Detección de ruido:** Mediante el módulo camera HD que incorpora el SDK del robot, se captura y procesa el flujo de audio, y se calculan los decibelios (dB). Se considera ruido cuando se superan los 75dB y se mantienen durante más de 1000 ms (para evitar confundir con otro tipo de sonidos como chirridos o golpes). En este caso, el robot reacciona al ruido quejándose del volumen o preguntando: "¿Qué es ese ruido?".

**Localización de sonido:** La localización se realiza mediante el módulo de control de hardware que incorpora el SDK del robot, el cual calcula el ángulo respecto al robot donde se ha detectado el sonido. Tras esta detección el robot reacciona mediante un movimiento de cabeza hacia la fuente de sonido.

**Detección de personas:** Esta detección se realiza mediante los sensores infrarrojos y la cámara HD que incorpora el robot en la parte frontal de su cabeza. En este segundo caso, el módulo gestor de multimedia del SDK procesa y envía la imagen a un servidor local, donde se encuentra desplegado un módulo de visión basado en inteligencia artificial desarrollado para agentes sociales interactivos (Ang et al., 2024). Este módulo procesa las imágenes recibidas y envía una respuesta al módulo reactivo, permitiendo así que el robot reaccione, en este caso mediante la reproducción de una frase.

**Ejecución de movimientos naturales:** Para los simular movimientos naturales se ha tenido en cuenta la movilidad del robot (movimiento de brazos en el eje vertical 270º, movimiento de cabeza eje vertical y horizontal 180º, movimiento de ruedas avanzar/retroceder y girar 360º) y se han programado movimientos de giro de cabeza, giro de torso y desplazamiento, los cuales se ejecutan a una velocidad y rango de movimiento reducidos para evitar la brusquedad de la acción.

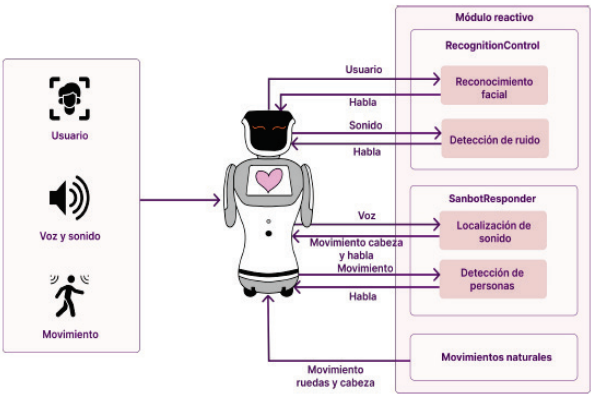



*Figura 5: Módulo reactivo.*

Además, el gestor de Consultas Externas (*External Queries*) permite consultar APIs externas y el gestor de Personalizaciones (*Customizations Managements*, que gestiona todos los parámetros de personalización del usuario y el robot.

### *3.4.3 Integración con otros componentes*

- Módulo de Inteligencia Artificial (*IA Module*): Procesa las solicitudes del usuario a través de peticiones HTTP, utilizando un modelo LLM para entender y responder adecuadamente. Este módulo integra dos componentes principales relacionados con los *endpoints* de la API de OpenAI, el primero es *ChatCompletions*, que permite interactuar con modelos de lenguaje basados en inteligencia artificial, para generar respuestas en conversaciones estructuradas. Para obtener la emoción del robot, se le solicita a *ChatCompletions* mediante el *prompt* que analice y categorice cada entrada del usuario según el modelo de Eckman y de una respuesta coherente en función de la emoción detectada intentando empatizar lo máximo posible con el usuario: no buscando imitarlo sino animarlo si está triste, calmarlo si está enfadado, etc. Asimismo, se solicita también que en la respuesta se tenga también en cuenta la edad y el nombre de la persona. El segundo componente es *AudioSpeech*, que permite la conversión de texto a voz y de voz a texto. Cada una con diferente tonalidad que, subjetivamente, se pueden considerar de un género concreto, además dicha API recibe un *prompt* opcionalmente, que nos permitió reducir la entonación anglosajona.

- Control del Robot (*Robot Control*): Utiliza las funcionalidades del SDK para gestionar acciones físicas como el movimiento de extremidades y la gestión del habla, así como el módulo de voz propio del robot.

## 4. Conversación natural y emocional entre una persona y el robot

Una vez presentado el sistema y explicados los distintos módulos que lo componen, se muestra en este apartado un ejemplo de flujo conversacional entre una persona y el robot. El proceso comienza cuando el usuario selecciona el botón que inicia la conversación:

**Respuesta del robot**: ¡Hola Lucía!, ¿cómo estás hoy?
El sistema ha podido reconocer al usuario, por lo que la ha llamado por su nombre.
**Entrada del usuario:** "¡Hola! Estoy bien, con muchas ganas de conversar"

El módulo conversacional recibe esta entrada, y se la envía a OpenAI mediante peticiones HTTP, generando la posible respuesta:
**Respuesta del robot:** [(1)/(1)] ¿Sobre qué te gustaría conversar? <Lucia,5>

Como se puede observar, la respuesta está compuesta por tres partes:

1. Interpretación emocional: Los números entre corchetes, como (1)/(1), representan emociones. El primer número (1) corresponde a la emoción del usuario, que en este caso sería "Alegría". El segundo número (1) es la emoción que el robot debe transmitir en su respuesta, en este caso sería "Alegría" igualmente. La emoción del usuario se utiliza para adaptar la conversación en el módulo conversacional, mientras que la emoción del robot se procesa en el módulo emocional para definir cómo debe expresarse físicamente.
2. Respuesta textual: Es el texto generado y verbalizado por el modelo, que en este caso es: "¿Sobre qué te gustaría conversar?"
3. Datos del usuario: En este apartado se incluye información personal proporcionada por el usuario o por el módulo reactivo, como su nombre y edad, en este caso: <Lucia,5>.

El Módulo Emocional transforma el dato sobre la emoción que debe transmitir el robot en acciones de los actuadores del robot, es decir en la expresión física de la emoción. Eso incluye expresión facial (ver Figura 2) y expresión corporal, en la forma de movimientos de brazos y cabeza, y activación de luces LED de diferentes colores. Es el Módulo Conversacional el que combina la respuesta textual con las respuestas físicas para comunicarse con el usuario. En el ejemplo dado, el robot respondería, mientras muestra una expresión facial de alegría y realiza movimientos que refuerzan esta emoción. Este flujo conversacional permite que el robot no solo responda de manera coherente, sino que también exprese emociones de manera natural. El Módulo Reactivo detecta en una primera instancia el rostro de la persona, para comprobar si se encuentra en su base de datos, en caso de que sea así, comienza a hablar mencionando su nombre y adapta la conversación según la edad del usuario. Los datos del usuario se utilizan para personalizar la conversación, ya que en momentos de inactividad el robot cada cierto tiempo le pregunta al usuario si desea continuar la conversación, llamándole por su nombre. Además, es importante manejar la edad como un dato esencial para que las respuestas del robot estén enfocadas al grupo generacional del usuario.
Por otra parte, el robot, gracias al Módulo Reactivo, siempre está pendiente a sonidos o movimientos de su entorno, pudiendo darse las siguientes situaciones:

**Ejecución de movimientos naturales:** Mientras Lucía mantiene la conversación con el robot este ejecuta pequeños movimientos de cabeza o de torso cada pocos segundos, lo que ayuda a que Lucía perciba el robot como un objeto más animado, y no tan rígido y robótico.

**Detección de ruido:** Si el nivel de ruido en la sala se mantiene por encima del umbral establecido, el sistema detecta durante la conversación que hay ruido, e interviene introduciendo una frase aleatoria como "Por favor, bajar el volumen". En ese caso, la interacción sería la siguiente:

**Robot:** ¡Hola Lucía!, ¿cómo estás hoy?
**Usuario:** "¡Hola! Estoy bien, con muchas ganas de conversar"
**Grupo de personas:** <<Barullo>>
**Robot:** "Por favor, bajar el volumen".
**Robot:** ¿Sobre qué te gustaría conversar?

**Localización de sonido:** Si durante la conversación con Lucía alguien da un portazo o se produce algún sonido detectable por el sistema, mediante la localización de sonido el robot girará su cabeza hacia la fuente de sonido y dirá una frase como: "Me parece haber escuchado algo por allí" y la conversación seguirá su curso en base a la respuesta de Lucía.

**Detección de personas:** Si durante la conversación el sistema detecta que alguien pasa por enfrente o por detrás del robot, éste emitirá una respuesta como: "¡Me asustaste!

No te había visto" y si la detección es por detrás, girará su torso. Por ejemplo:

**Robot:** ¡Hola Lucía!, ¿cómo estás hoy?
**<<Una persona pasa por detrás>>**
**Robot:** <<Gira torso>> "¡Me asustaste! No te había visto" <<Gira torso>>
**Usuario 1:** "Parece que te ha visto ¿Quieres saludar?"
**Usuario 2:** "Hola, ¿Cómo estás?"
**Respuesta del robot:** "Estoy bien, pero me había asustado."

**Reconocimiento facial:** Durante la conversación, el sistema es capaz de reconocer expresiones faciales, edad y género, y la persona con la que interacciona. Ante este reconocimiento el sistema reacciona mediante una frase que el robot introduce en la conversación, mencionando alguna de las características detectadas. En este caso el reconocimiento facial permite que se realice la identificación del usuario Lucía:

**Robot:** ¡Hola Lucía!, ¿cómo estás hoy? Te veo un poco triste.
**Usuario:** "Tienes razón, he perdido mi pulsera favorita y me ha dado mucha rabia"
**Respuesta del robot:** "Vaya, no te preocupes, seguro que está donde menos te lo esperas."

## 5. Evaluación con usuarios

El objetivo fundamental de la primera evaluación realizada consistió en comprobar el correcto funcionamiento del módulo, valorar su usabilidad, aceptación y el impacto en la interacción de los aspectos emocionales y de personalización. Asimismo, se consultó sobre el tipo de voz a usar poniendo especial énfasis en la consideración de la variable de género. He de mencionar que, en esta primera evaluación, el módulo reactivo estaba en desarrollo, por lo cual no se pudo evaluar en esta instancia. A continuación, se presenta la metodología que se llevó a cabo en la evaluación, y el análisis de los resultados obtenidos.

### 5.1 Metodología

En la evaluación participaron 18 usuarios en total, 7 hombres y 11 mujeres con edades entre 18 y 83 años (*M= 41.16, SD=22.95*). Las sesiones se llevaron a cabo en el laboratorio del grupo de investigación con una duración de aproximadamente 30 minutos por sesión, y con la siguiente estructura:

**Cuestionario inicial de caracterización:** utilizado para conocer los datos sobre el usuario participante, como su edad, capacidades sensoriales al utilizar dispositivos electrónicos o frecuencia de uso de la tecnología.

**Prueba de voces:** se mostraban varias voces a las personas, cinco de las voces eran sacadas del *endpoint Audio Speech* de OpenAI y la voz restante era la propia del robot Sanbot; tres voces eran masculinas y tres femeninas, y se preguntaba al usuario por su preferencia.

**Conversación con el robot:** Con el objetivo de medir el impacto de la componente emocional, se utilizó la técnica Testing A/B que consiste en dividir a los usuarios en dos grupos, cada grupo evaluaría una versión del sistema desarrollado (Kumar, 2019): el grupo A interactuó con el sistema que tenía capacidad de realizar consultas a OpenAI y que contaba con sus voces, además incluía el análisis emocional del usuario y la expresión emocional del robot mediante gestos, expresión facial o luces LED, mientras que el grupo B interactuó con el sistema que únicamente tenía la capacidad de conversar con el robot, pero con la voz original del robot Sanbot. El usuario mantenía dos conversaciones con el robot, la primera con un tema libre que elegía el usuario y la segunda con un tema cerrado que elegía el evaluador. El objeto de esa segunda conversación era explotar la capacidad de crear conversaciones sobre situaciones ficticias en las que se pudieran encontrar usuario y robot (safari, viaje espacial, descubrimiento de un planeta…).

**Cuestionarios:** Preguntas específicas sobre el funcionamiento del módulo conversacional (ver Tabla 2). Cuestionario de usabilidad con siete preguntas extraídas del SUS (Jordan, 1996) (ver Tabla 3). Cuestionario de aceptación POST VAVA-Q (Cerezo et al., 2025), generado y validado en el grupo de investigación para evaluar la aceptación tecnológica (ver Tabla 4), y que aborda cuatro dimensiones diferentes: Control Percibido, Intención de Uso, Actitud hacia el uso y Norma Subjetiva.

*Tabla 2. Cuestionario de Preguntas de interacción por voz. Respuestas del tipo Likert (1-Totalmente en desacuerdo a 5-Totalmente de acuerdo)*

| Preguntas del cuestionario | |
|---|---|
| Preguntas interacción voz | He entendido bien lo que me decía el robot. |
| | Me pareció que al robot le costaba entender lo que le decía. |
| | Considero que la aparición de la respuesta en la pantalla me ha ayudado durante la conversación. |
| | Considero que la conversación que he tenido con el robot ha seguido un hilo coherente. |
| | Considero que el robot ha entendido correctamente mis consultas y ha sabido darles una respuesta con sentido. |
| | Considero que la voz del robot es agradable. |
| | Considero que la voz del robot suena artificial. |
| | Considero que la voz del robot es fácil de entender. |
| | Considero que la respuesta del robot es rápida. |

*Tabla 3. Cuestionario de Usabilidad. Respuestas del tipo Likert (1-Totalmente en desacuerdo a 5-Totalmente de acuerdo)*

| Preguntas del cuestionario | |
|---|---|
| Preguntas SUS | Creo que me gustaría utilizar este robot con frecuencia. |
| | Creo que necesitaría el apoyo de otra persona para poder utilizar este robot. |
| | Me imagino que la mayoría de la gente aprendería a utilizar este robot muy rápidamente. |
| | Encontré el robot innecesariamente complicado de usar. |
| | Me sentí muy seguro usando el robot. |
| | Encuentro el uso del robot raro. |
| | Siento que tengo que aprender muchas cosas antes de empezar a usar el robot. |

*Tabla 4. Cuestionario POST VAVA-Q. Respuestas del tipo Likert*

| Preguntas del cuestionario | Respuestas |
|---|---|
| Creo que con los recursos y conocimientos que tengo, seguir las instrucciones para usar el robot Sanbot Elf, me ha sido... | (1-Fácil a 5-Difícil) |
| Si se me propusiera otra vez, tengo la intención de volver a utilizar el robot Sanbot Elf en los próximos días. | (1-Falso a 5-Verdadero) |
| Valorando lo que me puede aportar utilizar el robot Sanbot Elf, creo que es una idea… | (1-Inútil a 5-Útil) |
| La posibilidad de contarles a las personas importantes para mí que estoy utilizando el robot Sanbot Elf, es una idea que... | (1-Me gusta a 5-No me gusta) |
| Al utilizar el robot Sanbot Elf, resolver los retos que se han presentado ha sido... | (1-Fácil a 5-Difícil) |
| Quiero volver a utilizar el robot Sanbot Elf en los próximos días. | (1-Falso a 5-Verdadero) |
| Utilizar el robot Sanbot Elf, ha sido una idea… | (1-Aburrida a 5-Divertida) |
| Si las personas que son importantes para mí supieran que estoy utilizando el robot Sanbot Elf, yo creo que… | (1-Les gustaría a 5-No les gustaría) |
| Yo creo que si me lo proponen, encontraré el tiempo para utilizar de nuevo el robot Sanbot Elf en los próximos días. | (1-Falso a 5-Verdadero) |

## 5.2 Resultados

A continuación, se muestran los resultados obtenidos de esta primera evaluación con usuarios, así como el análisis de los mismos y las mejoras a realizar. Durante la prueba de **Preferencia de Voces** se pudo comprobar que el 72.22% de las personas eligieron una voz que coincidía en género con el propio, aunque un mayor número de hombres (el 57% de ellos) y de mujeres (el 91% de ellas) eligieron una voz femenina (ver figura 6).

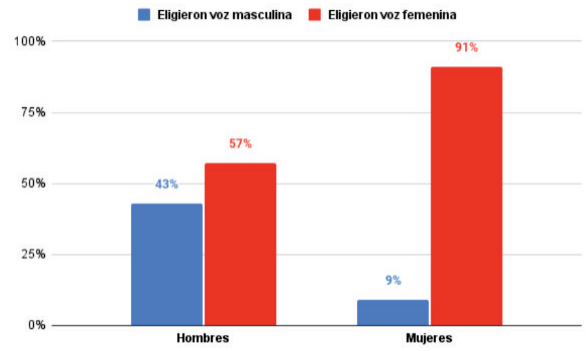



*Figura 6: Resultados de la prueba de preferencia género de voz.*

En cuanto a la **Usabilidad** (Preguntas SUS) los resultados fueron muy satisfactorios, pudiendo destacar las puntuaciones a preguntas como el uso frecuente del robot o la poca complicación del módulo conversacional. En el **análisis de género** (ver figura 7)**,** se pudo observar que existen algunas diferencias: las mujeres tienen más interés en usar el robot, mientras que los hombres califican en mayor medida el uso del robot como raro, pero valoran su aprendizaje a la hora de usarlo como más rápido.

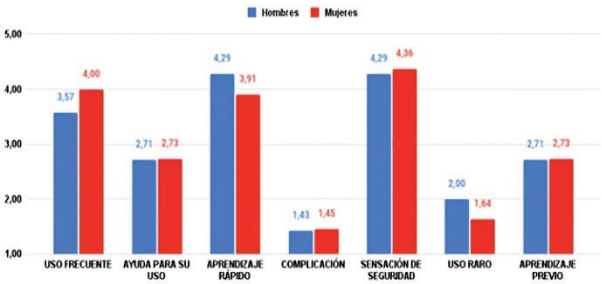



*Figura 7: Resultados de la prueba de usabilidad preguntas SUS.*

En cuanto al **análisis por grupo (A/B)**, el grupo A considera que necesitarían más ayuda para utilizar el robot, que su uso es más complicado y raro y que necesitan aprender muchas cosas antes de iniciar el módulo conversacional (ver Figura 8). Se cree que esto puede ser debido a la necesidad de pasar por las pantallas de personalización tanto del usuario como del robot antes de poder llevar a cabo la conversación con el robot. Ello pone de manifiesto la necesidad de mantener las interfaces sencillas al aumentar la complejidad del procesamiento.

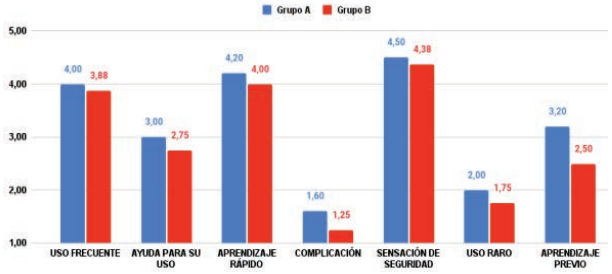



*Figura 8: Resultados de la prueba de usabilidad por grupo.*

También se analizó la **Usabilidad (Preguntas interacción por voz),** donde el resultado del análisis general arrojó que todos están de acuerdo o muy de acuerdo en que hubo una buena comprensión y entendimiento de la voz, tanto por parte del usuario como del robot, la coherencia de las conversaciones, la respuesta rápida del módulo o la aparición de la respuesta en pantalla. El grupo B ha comprendido mejor las respuestas del robot con un valor medio de 4.80, por lo que la incomprensión por parte del robot tuvo solo una media de1.75, probablemente debido a los problemas de pronunciación de las voces OpenAI (nativamente inglesas), mientras que el grupo A tuvo resultados un poco menos positivos, con medias de 4.80 y 2.00, respectivamente. Los participantes del grupo B opinaron que el robot entendía peor las consultas con una media de 4.75 respecto a la media del grupo A (M=4.90) debido a que en ese grupo no se aplicaba ninguna personalización. Por último, el grupo A (M=4.20) consideró las respuestas del robot más lentas que las del grupo B (M=4.75), seguramente debido a que la petición a OpenAI contenía más datos de tipo emocional, personalización o similares. Finalmente, se analizó la **Aceptación**, donde los resultados fueron muy satisfactorios, pudiendo observar que todos los usuarios consideraron que el sistema no era difícil de utilizar a pesar de la existencia de diversas edades o experiencias con la tecnología, y que muchos de ellos tenían intención de usarlo nuevamente por su utilidad en base a sus necesidades, solo dos votaron de forma neutral. También se pudo observar que las mujeres encuentran más útil el uso del robot (M=4.91) respecto a los hombres (M=4.29), sin embargo, los hombres opinaron que tienen mayor intención de usarlo nuevamente (M=4.86) respecto a las mujeres (M=4.55). En general, consideraron que la prueba fue muy divertida.

Analizando los resultados en su conjunto, se puede concluir que todos los participantes calificaron el uso del robot como fácil y la prueba como divertida. Todos ellos estuvieron muy de acuerdo o de acuerdo con que entendieron bien lo que les decía el robot y que la conversación siguió un hilo coherente, de forma que el robot había entendido la pregunta y había sabido contestar adecuadamente. Además, encontraron de gran ayuda el que la respuesta se mostrara por pantalla. Si bien aquellos que probaron el modelo emocional completo opinaron que el robot entendía mejor sus consultas, calificaron al robot como más lento y difícil de usar. El segundo punto puede estar relacionado con la necesidad de pasar, por varias pantallas de personalización, tanto del robot como del usuario, por lo que habría que reconsiderarlas o rediseñarlas.

### 5.3 Mejoras Realizadas

Debido a los resultados antes planteados, se realizaron diversas modificaciones para mejorar el sistema conversacional desarrollado. En primer lugar, se modificó la vista inicial de configuración para que sea más sencilla e intuitiva y el usuario no necesite navegar por otras pantallas de la aplicación, seleccionando "Ajuste Automático" el usuario ya no necesita insertar su nombre y edad, ya que esta funcionalidad le redirige a la pantalla de conversación, donde el robot lo identificará mediante el reconocimiento de rostro del módulo reactivo, en caso de que no lo reconozca le preguntará su nombre y edad para la personalización de la experiencia. En segundo lugar, la generación de voz se realizó mediante el uso de *prompts* adaptativos, donde se le indicaba que debía ser hispanohablante para así obtener una entonación más natural y con menos acento anglosajón, además se situó en la pantalla inicial de configuración las cuatro voces con menor acento foráneo (ver Figura 9).


*Figura 9: Pantalla de configuración modificada*

Además, se ha optimizado el código para que el tiempo de procesamiento de los diferentes módulos sean mínimos e imperceptibles para el usuario. Sin embargo, el tiempo de

respuesta de los *endpoints* de OpenAI dependerá de la velocidad y latencia de la red que está configurada en el robot. Con estas mejoras implementadas se realizó una primera evaluación muy preliminar en una residencia de personas mayores. De manera empírica se pudo observar que la mayoría de las personas usuarias mantuvieron una conversación fluida y natural con el robot, y resultó interesante comprobar que algunos participantes compartían con los robots aspectos personales. En general tanto el robot como la aplicación tuvieron una buena aceptación por parte de los usuarios. También se realizó otra prueba informal durante un taller de co-creación con adultos y personas mayores. En este caso, la experiencia no fue tan satisfactoria debido al elevado nivel de ruido ambiental, que activaba con frecuencia el módulo reactivo del robot. Esta situación evidenció que, en entornos grupales o ruidosos, es necesario ajustar la sensibilidad del sistema para evitar reacciones no deseadas. En términos generales, la aceptación del robot fue muy positiva y la latencia casi imperceptible. Sin embargo, es necesario hacer un estudio más exhaustivo utilizando instrumentos cuantitativos que permitan evaluar en detalle las percepciones de los usuarios. De este modo será posible determinar qué diferencias hay entre las experiencias realizadas en el laboratorio y las realizadas en estos contextos no controlados.

## 6. Conclusiones y trabajo futuro

En este trabajo se ha presentado un sistema conversacional diseñado para fortalecer la interacción natural entre personas y robots, incorporando inteligencia artificial para generar diálogos fluidos acompañados de expresiones faciales, movimientos y colores que permiten reflejar las emociones del robot, y reaccionar de forma natural en base a situaciones que detecta el robot en su entorno. El sistema desarrollado fue evaluado, de forma preliminar, con 18 usuarios, evidenciando resultados positivos en términos de usabilidad y aceptación, aunque también se identificaron algunas limitaciones, como la utilización de voces no idóneas para la interacción en español, el incremento en los tiempos de procesamiento al integrar componentes emocionales y la necesidad de mantener interfaces de configuración simples frente a la complejidad de los módulos de personalización. A partir de estos resultados, se llevaron a cabo modificaciones en el desarrollo del sistema para solucionar los problemas detectados. Sin embargo, aunque se realizaron algunas evaluaciones empíricas, aún es necesario evaluar la aplicación, de forma más exhaustiva y rigurosa. Por ello, como trabajo futuro a corto plazo, se debe tener en cuenta, reducir la latencia o fallos que la aplicación pudiera tener en relación con la conexión a internet en entornos externos al laboratorio, para esto se prevé implantar un modelo LLM en un servidor local que se encuentre en la misma red del centro en el que se realicen las pruebas y se pretende llevar a cabo una segunda fase de evaluación con el fin de analizar el impacto de estos nuevos componentes en la experiencia de interacción.

Esta fase incluirá otros instrumentos de evaluación como *GoodSpeed* (Bartneck, Croft y Kulic, 2009) o *RoSAS* (Pan, Croft, y Niemeyer, 2018) teniendo en cuenta también un mayor número y diversidad de participantes y la aplicación de protocolos estandarizados en el ámbito de la interacción humano-robot (HRI), así como el uso de instrumentos específicos para medir variables como la empatía percibida, la fluidez conversacional y la naturalidad en la comunicación. Por otra parte, también sería interesante evaluar si las expresiones faciales, los movimientos y los colores que muestra el robot para demostrar sus emociones, tienen algún impacto en la percepción por parte de los usuarios, de forma similar al análisis realizado por (Raggioli et al., 2025) para determinar si ciertas expresiones y movimientos aumentaban la tasa de reconocimiento de ciertas emociones. De este modo, se podrán validar las mejoras implementadas, así como detectar nuevas necesidades de los usuarios y profundizar en el estudio de la interacción humano-robot.

## Agradecimientos

Este trabajo está parcialmente financiado por el Ministerio de Ciencia e Innovación y Universidades (MCIU), la Agencia Estatal de Investigación (AEI) y la UE (FEDER) a través del contrato PID2022-136779OB-C31, por MCIN/ AEI/10.13039/501100011033/ y por la Unión Europea NextGenerationEU/PRTR a través del proyecto TED2021-130374B-C22 y por el Gobierno de Aragón (Grupo T60_23R).

## Referencias

Ang E., Bejleri A., Tantisira B, Van de Velde. A., (2024), Considerations for the future of social robots and human-robot interactions. URL: https://www.oxjournal.org/the-future-of-social-robots-and-human-robot-interactions/.

Alnajjar, F., Khalid, S., Vogan, A. A., Shimoda, S., Nouchi, R., & Kawashima, R. (2019). Emerging cognitive intervention technologies to meet the needs of an aging population: a systematic review. Frontiers in Aging Neuroscience, 11, 291.

Assad-Uz-Zaman, M., Rasedul Islam, M., Miah, S., & Rahman, M. H. (2019). NAO robot for cooperative rehabilitation training. Journal of rehabilitation and assistive technologies engineering, 6, 2055668319862151.

Babel, F., Kraus, J., Miller, L., Kraus, M., Wagner, N., Minker, W., & Baumann, M. (2021). Small talk with a robot? The impact of dialog content, talk initiative, and gaze behavior of a social robot on trust, acceptance, and proximity. International Journal of Social Robotics, 13(6), 1485-1498.

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. International journal of social robotics, 1(1), 71-81.

Bonarini, A. (2020). Communication in human-robot interaction. Current Robotics Reports, 1(4), 279-285.

Brave, S., Nass, C., & Hutchinson, K. (2005). Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent. International journal of human-computer studies, 62(2), 161-178.

Breazeal, C. (2003). Emotion and sociable humanoid robots. International journal of human-computer studies, 59(1-2), 119-155.

Calvo-Barajas, N., Perugia, G., & Castellano, G. (2020, August). The effects of robot's facial expressions on children's first impressions of trustworthiness. In 2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN) (pp. 165-171). IEEE.

Cerezo, E., Lacuesta, R., Gallardo, J., & Aguelo, A. (2025). Exploring the use of voice assistants in nursing homes. International Journal of Human–Computer Interaction, 1-17.

Chen, Y., Garcia-Vergara, S., & Howard, A. M. (2018). Effect of feedback from a socially interactive humanoid robot on reaching kinematics in children with and without cerebral palsy: a pilot study. Developmental neurorehabilitation, 21(8), 490-496.

Cherakara, N., Varghese, F., Shabana, S., Nelson, N., Karukayil, A., Kulothungan, R., & Lemon, O. (2023). Furchat: An embodied conversational agent using llms, combining open and closed-domain dialogue with facial expressions. arXiv preprint arXiv:2308.15214.

Churamani, N., Kalkan, S., & Gunes, H. (2020, August). Continual learning for affective robotics: Why, what and how?. In 2020 29th IEEE international conference on robot and human interactive communication (RO-MAN) (pp. 425-431). IEEE.

Cross, E. S., Hortensius, R., & Wykowska, A. (2019). From social brains to social robots: applying neurocognitive insights to human–robot interaction. Philosophical Transactions of the Royal Society B, 374(1771), 20180024.

de Graaf, M. M., Ben Allouch, S., & Van Dijk, J. A. (2015, October). What makes robots social?: A user's perspective on characteristics for social human-robot interaction. In International Conference on Social Robotics (pp. 184-193). Cham: Springer International Publishing.

Dziergwa, M., Kaczmarek, M., Kaczmarek, P., Kędzierski, J., & Wadas-Szydłowska, K. (2018). Long-term cohabitation with a social robot: a case study of the influence of human attachment patterns. International Journal of Social Robotics, 10(1), 163-176.

Ekman, P. (2014). Expression and the nature of emotion. Approaches to emotion, 319-343.

Gou, M. S., Vouloutsi, V., Grechuta, K., Lallée, S., & Verschure, P. F. (2014, July). Empathy in humanoid robots. In Conference on Biomimetic and Biohybrid Systems (pp. 423-426). Cham: Springer International Publishing.

Henschel, A., Laban, G., & Cross, E. S. (2021). What makes a robot social? A review of social robots from science fiction to a home or hospital near you. Current Robotics Reports, 2(1), 9-19.

Jiang, Y., Shao, S., Dai, Y., & Hirota, K. (2024, July). A LLM-Based Robot Partner with Multi-modal Emotion Recognition. In International Conference on Intelligent Robotics and Applications (pp. 71-83). Singapore: Springer Nature Singapore.

Jordan, P. W., Thomas, B., McClelland, I. L., & Weerdmeester, B. (Eds.). (1996). Usability evaluation in industry. CRC press.

Kim, C. Y., Lee, C. P., & Mutlu, B. (2024, March). Understanding large-language model (llm)-powered human-robot interaction. In Proceedings of the 2024 ACM/IEEE international conference on human-robot interaction (pp. 371-380).

Kumar, R. (2019, March). Data-driven design: Beyond a/b testing. In Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (pp. 1-2).

Kyprianou, G., Karousou, A., Makris, N., Sarafis, I., Amanatiadis, A., & Chatzichristofis, S. A. (2023). Engaging learners in educational robotics: Uncovering students' expectations for an ideal robotic platform. Electronics, 12(13), 2865.

Laban, G., & Cross, E. S. (2024). Sharing our Emotions with Robots: Why do we do it and how does it make us feel?. IEEE Transactions on Affective Computing.

Lee, M. K., Forlizzi, J., Kiesler, S., Rybski, P., Antanitis, J., & Savetsila, S. (2012, March). Personalization in HRI: A longitudinal field experiment. In Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction (pp. 319-326).

Leite, I., Pereira, A., Mascarenhas, S., Martinho, C., Prada, R., & Paiva, A. (2013). The influence of empathy in human–robot relations. International journal of human-computer studies, 71(3), 250-260.

Melo, F., & Moreno, P. (2022, April). Socially reactive navigation models for mobile robots. In 2022 IEEE international conference on autonomous robot systems and competitions (ICARSC) (pp. 91-97). IEEE.

Mohebbi, A. (2020). Human-robot interaction in rehabilitation and assistance: a review. Current Robotics Reports, 1(3), 131-144.

Paiva, A., Leite, I., Boukricha, H., & Wachsmuth, I. (2017). Empathy in virtual agents and robots: A survey. ACM Transactions on Interactive Intelligent Systems (TiiS), 7(3), 1-40.

Pan, M. K., Croft, E. A., & Niemeyer, G. (2018, February). Evaluating social perception of human-to-robot handovers using the robot social attributes scale (rosas). In Proceedings of the 2018 ACM/IEEE international conference on human-robot interaction (pp. 443-451).

Pinto-Bernal, M., Biondina, M., & Belpaeme, T. (2025). Designing Social Robots with LLMs for Engaging Human Interaction. Applied Sciences, 15(11), 6377.

Raggioli, L., Esposito, R., Rossi, A., & Rossi, S. (2025). Exploring the Role of Robot's Movements for a Transparent Affective Communication. IEEE Robotics and Automation Letters.

Rawal, N., Maharjan, R. S., Romeo, M., Bigazzi, R., Baraldi, L., Cucchiara, R., & Cangelosi, A. (2024, September). Intelligent multimodal artificial agents that talk and express emotions. In International Workshop on Human-Friendly Robotics (pp. 240-254). Cham: Springer Nature Switzerland.

Reimann, M. M., Kunneman, F. A., Oertel, C., & Hindriks, K. V. (2024). A survey on dialogue management in human-robot interaction. ACM Transactions on Human-Robot Interaction, 13(2), 1-22.

Robinson, N. L., Cottier, T. V., & Kavanagh, D. J. (2019). Psychosocial health interventions by social robots: systematic review of randomized controlled trials. Journal of medical Internet research, 21(5), e13203.

Romat, H., Williams, M. A., Wang, X., Johnston, B., & Bard, H. (2016, March). Natural human-robot interaction using social cues. In 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI) (pp. 503-504). IEEE.

Sarrica, M., Brondi, S., & Fortunati, L. (2020). How many facets does a "social robot" have? A review of scientific and popular definitions online. Information Technology & People, 33(1), 1-21.

Scoglio, A. A., Reilly, E. D., Gorman, J. A., & Drebing, C. E. (2019). Use of social robots in mental health and well-being research: systematic review. Journal of medical Internet research, 21(7), e13322.

Son, E. (2022). Visual, Auditory, and Psychological Elements of the Characters and Images in the Scenes of the Animated Film, Inside Out. Quarterly Review of Film and Video, 39(1), 225-240.

Spezialetti, M., Placidi, G., & Rossi, S. (2020). Emotion recognition for human-robot interaction: Recent advances and future perspectives. Frontiers in Robotics and AI, 7, 532279.

Van Oost, E., & Reed, D. (2010, June). Towards a sociological understanding of robots as companions. In International conference on human-robot personal relationship (pp. 11-18). Berlin, Heidelberg: Springer Berlin Heidelberg.

Wang, C., Hasler, S., Tanneberg, D., Ocker, F., Joublin, F., Ceravola, A., & Gienger, M. (2024, May). Lami: Large language models for multi-modal human-robot interaction. In Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (pp. 1-10).

Wullenkord, R., & Eyssel, F. (2020). Societal and ethical issues in HRI. Current Robotics Reports, 1(3), 85-96.

Xu, J., Broekens, J., Hindriks, K., & Neerincx, M. A. (2015). Mood contagion of robot body language in human robot interaction. Autonomous Agents and Multi-Agent Systems, 29(6), 1216-1248.

Yang, G. Z., J. Nelson, B., Murphy, R. R., Choset, H., Christensen, H., H. Collins, S., ... & McNutt, M. (2020). Combating COVID-19—The role of robotics in managing public health and infectious diseases. Science robotics, 5(40), eabb5589.

INTERACCION
Revista Digital de AIPO
Asociación Interacción Persona-Ordenador
Vol. 6, No 2 (2025)
ISSN 2695-6578

# UIBAIFED: Un dataset de expresiones faciales generado por IA para visibilizar la diversidad

## UIBAIFED: An AI-generated facial expression dataset to enhance diversity

**Esperança Amengual-Alcover**
Departament de Ciències Matemàtiques i informàtica
Universitat de les Illes Balears
Palma de Mallorca, España
eamengual@uib.es

**Maria Francesca Roig-Maimó**
Departament de Ciències Matemàtiques i informàtica
Universitat de les Illes Balears
Palma de Mallorca, España
xisca.roig@uib.es

**Ramon Mas-Sansó**
Departament de Ciències Matemàtiques i informàtica
Universitat de les Illes Balears
Palma de Mallorca, España
ramon.mas@uib.es

**Miquel Mascaró-Oliver**
Departament de Ciències Matemàtiques i informàtica
Universitat de les Illes Balears
Palma de Mallorca, España
miquel.mascaro@uib.cat

Recibido: 15.11.2025 | Aceptado: 09.12.2025

**Palabras Clave**

Interacción persona-ordenador
IPO
Aprendizaje profundo
Conjunto de datos de expresiones faciales
FER

**Resumen**

Este artículo presenta UIBAIFED, un nuevo conjunto de datos de expresiones faciales compuesto por imágenes realistas y de alta calidad, etiquetadas con grupo de edad, género, etnia y 22 microexpresiones basadas en las expresiones universales de Ekman y la taxonomía de Gary Faigin. El conjunto de datos fue generado mediante modelos de difusión y está diseñado para mejorar la investigación en reconocimiento de expresiones faciales, aumentado la representación étnica, generacional y de complexión física, visibilizando la diversidad existente en el mundo real. La validación inicial mediante una red neuronal convolucional (CNN) alcanzó una precisión del 80% en la clasificación de microexpresiones. Además, se realizó un experimento con participantes humanos con un subconjunto de datos formado por imágenes del grupo de edad de personas mayores (85 años), colectivo tradicionalmente marginado en los conjuntos de datos existentes. Los mejores resultados de clasificación fueron para las emociones de tristeza (94.6%), neutral (92.8%) y alegría (83.8%), mientras que las peores emociones identificadas fueron las emociones de miedo (33.9%) y asco (34.7%), con una precisión global del 74%. UIBAIFED ofrece un nivel de etiquetado más detallado que los conjuntos de datos existentes, lo que facilita el análisis del rendimiento en sistemas de reconocimiento de expresiones faciales en distintos grupos demográficos y contribuye al desarrollo de modelos más robustos y generalizables.

**Keywords**

Human Computer Interaction
HCI
Machine learning
Facial expression dataset
FER

**Abstract**

This article presents UIBAIFED, a new facial expression dataset composed of high-quality, realistic images labelled with age group, ethnicity, and 22 micro-expressions based on Ekman's universal expressions and Gary Faigin's taxonomy. The dataset was generated using diffusion models and is designed to support advanced research in Facial Expression Recognitions (FER), increasing ethnical, generational and complexity representativity, enhancing real world diversity. The initial validation using a Convolutional Neural Network (CNN) achieved an accuracy of 80% in micro-expression classification. To complement this automatic evaluation, an experiment with human participants was conducted using a subset of data composed by images corresponding to elder adults (85 years old), a group traditionally underrepresented in existing datasets. The best results were obtained when classifying sadness (94.6%), neutral (92.8%) and joy (83.8%) emotions. We got the worst results when classifying fear (33.9%) and disgust (34.7%) emotions. The overall accuracy was of 74%. UIBAIFED provides a more detailed annotation level than existing facial expression datasets, enabling performance analysis of FER systems across different demographic groups and contributing to the development of more robust and generalizable models.

## 1. Introducción

El reconocimiento de expresiones faciales (FER, por sus siglas en inglés *Facial Expression Recognition*) ha experimentado avances significativos en los últimos años, impulsados en gran medida por las mejoras técnicas de aprendizaje profundo (Li & Weng, 2022) y la creciente disponibilidad de conjuntos de datos (*datasets*) de alta calidad (Kollias et al., 2020). Estos conjuntos de datos desempeñan un papel crucial en el entrenamiento de modelos capaces de interpretar con precisión expresiones faciales en diferentes contextos. Sin embargo, los conjuntos de datos existentes aún presentan desafíos relacionados con la diversidad demográfica, los desequilibrios entre clases y cuestiones éticas como el sesgo en la representación (Buolamwini & Gebru, 2018). Abordar estos problemas es esencial para desarrollar modelos de FER más robustos y generalizables.

A pesar de la creciente disponibilidad de conjuntos de datos para el reconocimiento de expresiones faciales, *datasets* ampliamente utilizados como Fer2013 (Goodfellow et al., 2013), CK+ (Lucey et al., 2010), RAF-DB (Li et al., 2017) y AffectNet (Mollahosseini et al., 2019) presentan limitaciones. Entre ellas, se incluyen desequilibrios entre clases, donde algunas emociones, como por ejemplo la felicidad, están sobrerrepresentadas, mientras que otras, como el miedo o el asco, permanecen infrarrepresentadas (Fan et al., 2022). Además, muchos conjuntos de datos están compuestos principalmente por poblaciones jóvenes y de origen occidental, lo que limita la capacidad de generalización de los modelos de FER hacia grupos menos representados (Domínguez-Catena et al., 2024).

Para mitigar estas limitaciones, investigaciones previas han explorado enfoques alternativos, como las técnicas de aumento de datos (*data augmentation*) y la generación sintética de expresiones faciales, con el fin de mejorar la diversidad de los datos y el rendimiento de los modelos (Psaroudakis & Kollias, 2022). Sin embargo, hasta donde llega nuestro conocimiento, no existe actualmente ningún conjunto de datos de FER disponible públicamente que haya sido generado íntegramente mediante Inteligencia Artificial (IA).

En este trabajo presentamos UIBAIFED *(UIB Artificial Intelligence Facial Expression Dataset*), el primer conjunto de datos generado mediante inteligencia artificial diseñado para mejorar el entrenamiento y la evaluación de los modelos de reconocimiento de expresiones faciales. A diferencia de los conjuntos de datos tradicionales, UIBAIFED utiliza técnicas de inteligencia artificial generativa para crear un conjunto diverso y equilibrado de expresiones faciales. Este enfoque garantiza un corpus de entrenamiento más representativo para los sistemas modernos de FER, reduciendo los sesgos demográficos y mejorando la robustez general de los modelos.

## 2. Trabajo relacionado

### 2.1 Conjuntos de datos tradicionales de FER

Varios conjuntos de datos han sido ampliamente utilizados en la investigación sobre reconocimiento de expresiones faciales, entre ellos *FER2013*, CK+, RAF-DB y *AffectNet*. Estos conjuntos de datos han contribuido de manera significativa al avance de los modelos de aprendizaje profundo para la clasificación de emociones; sin embargo, con frecuencia presentan limitaciones tales como:

1. **Desequilibrios demográficos**: muchos conjuntos de datos se centran en poblaciones jóvenes y de origen occidental, lo que da lugar a modelos que generalizan de forma deficiente en grupos menos representados (Domínguez-Catena et al., 2024).
2. **Desequilibrios entre clases**: algunas emociones, como la felicidad y la emoción neutra, están representadas con mayor frecuencia que otras, como el miedo o el asco, lo que puede generar un rendimiento sesgado en los modelos (Zhang et al., 2023).
3. **Inconsistencias en el etiquetado**: las diferencias en la forma en que se anotan las emociones entre distintos conjuntos de datos pueden introducir ruido y dificultar la capacidad de generalización de los modelos (Chen & Joo, 2021).

Estas limitaciones han motivado a los investigadores a desarrollar nuevos conjuntos de datos más equilibrados y diversos que permitan una mejor capacidad de generalización de los modelos de FER.

### 2.2 Reconocimiento de microexpresiones

Las microexpresiones son expresiones faciales breves e involuntarias que revelan emociones genuinas. Su naturaleza efímera las hace difíciles de capturar y clasificar, aunque resultan fundamentales en campos como la psicología, la seguridad y la interacción persona-ordenador (Adegun & Vadapalli, 2020).

Una de las principales carencias de los conjuntos de datos actuales de FER es la ausencia de un etiquetado sistemático de las microexpresiones. A diferencia de los conjuntos de datos estándar, que se centran en categorías emocionales más amplias, las microexpresiones requieren una mayor granularidad y una anotación precisa. Esta limitación dificulta

el desarrollo de modelos capaces de detectar indicios emocionales sutiles en aplicaciones en tiempo real (Guerdelli et al., 2022).

La categorización de expresiones faciales propuesta por Faigin ofrece un marco integral para comprender la dinámica facial más allá de las siete categorías emocionales tradicionales (Faigin, 1990). Esta taxonomía destaca la complejidad de las expresiones, capturando variaciones sutiles que se suelen pasar por alto en los estudios convencionales de FER. Sin embargo, los conjuntos de datos existentes rara vez incorporan este nivel de detalle, lo que limita la capacidad de los modelos actuales para reconocer los matices emocionales. Superar esta brecha requiere conjuntos de datos diseñados explícitamente para alinearse con la categorización de Faigin.

## 3. Métodos

Para abordar los desafíos mencionados anteriormente, en este trabajo presentamos UIBAIFED, un conjunto de datos generado mediante inteligencia artificial, diseñado para ofrecer una presentación más equilibrada y diversa de las expresiones faciales. Al aprovechar el potencial de los modelos generativos, garantizamos variaciones controladas en cuanto a edad, género y etnia, manteniendo al mismo tiempo diferencias realistas en la postura, la iluminación y la intensidad de la expresión. Este enfoque tiene como objetivo mitigar los sesgos presentes en los conjuntos de datos tradicionales y mejorar la robustez de los modelos de reconocimiento de expresiones faciales.

### 3.1 Modelos faciales

Para garantizar la calidad del conjunto de datos, las imágenes generadas cumplen con los siguientes criterios: el rostro debe estar centrado y ocupar entre el 40% y el 70% del área de la imagen; la iluminación debe ser suficiente para resaltar con claridad los detalles de la expresión facial, mientras que el fondo debe permanecer uniforme y neutro, a fin de evitar posibles interferencias en la clasificación. Además, las expresiones faciales deben replicar con precisión las descripciones propuestas por Gary Faigin (1990). Asimismo, los defectos visuales deben ser mínimos y no comprometer la expresividad del rostro.

El conjunto de datos UIBAIFED garantiza una distribución equilibrada en cuanto al sexo, cinco grupos de edad diferenciados (véase la figura 1) y tres categorías de composición corporal (véase la figura 2).


*Figura 2: Distribución de grupos de edad en el conjunto de datos UIBAIFED.*

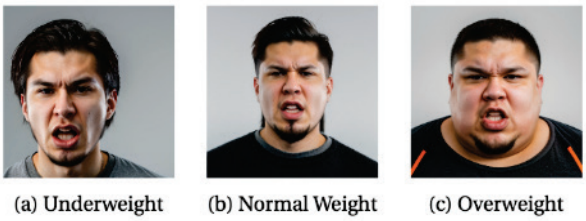



*Figura 2: Categorías de composición corporal en el conjunto de datos UIBAIFED.*

La diversidad étnica se considera según la clasificación establecida por la OMB (*Office of Management and Budget*) de los Estados Unidos (Office of Institutional Research, s.f.), la cual incluye grupos como nativos americanos, asiáticos, personas negras, hispanos, nativos de Hawái u otras islas del Pacífico y personas blancas de ascendencia europea, norteafricana o de Oriente Medio (véase la figura 3).

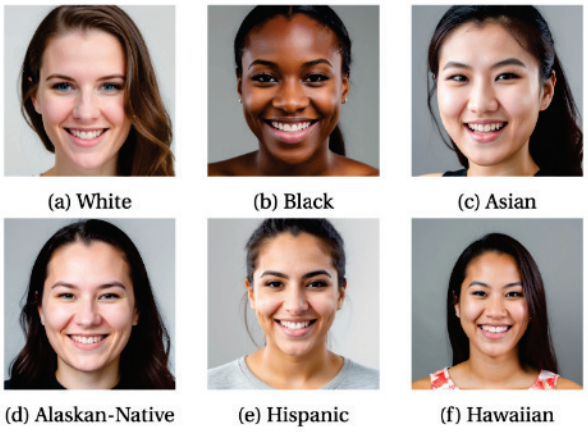



*Figura 3: Distribución de la diversidad étnica en el conjunto de datos UIBAIFED.*

### 3.2 Proceso de generación y filtrado de imágenes

Para la generación de imágenes de expresiones faciales en el conjunto de datos UIBAIFED, se utilizó el modelo *Stable Diffusion* (Stable Diffusion AI, s.f.) Esta tecnología de código abierto puede ejecutarse localmente, lo que ofrece la ventaja de generar un número ilimitado de imágenes. Su naturaleza flexible y las continuas contribuciones de la comunidad han permitido el desarrollo de versiones mejoradas, aumentando la variedad y la calidad de los resultados y garantizando que las imágenes cumplan con los criterios establecidos para el análisis de expresiones faciales.

Los *checkpoints* de *Stable Diffusion* son modelos pre-entrenados diseñados para generar imágenes a partir de descripciones textuales, donde se emplean grandes conjuntos de datos para aprender las correlaciones entre palabras y elementos visuales. La selección de un *checkpoint* requiere considerar su capacidad para generar una amplia variedad de imágenes con todas las características requeridas, al mismo tiempo que se minimiza el tiempo de generación. En base a hallazgos empíricos, se determinó que el *checkpoint Realistic Vision* (Civitai, s.f.-b) es el que mejor se adapta a las necesidades del conjunto de datos.

Para optimizar el modelo en la generación de expresiones faciales, se empleó la técnica *Low-Rank Adaptation (LoRA)* (Hu et al., 2022; Zeng and Lee, 2023). *LoRA* permite adaptar rápidamente modelos de aprendizaje automático a nuevos contextos mediante la incorporación de componentes ligeros al modelo original, en lugar de modificar toda su estructura. En el caso de *Stable Diffusion*, se utilizaron *LoRAs* específicamente entrenadas para la generación de expresiones faciales, obtenidas de *CivitAI* (Civitai, s.f.-a). La tabla 1 muestra las *LoRAs* utilizadas para la generación del conjunto de datos UIBAIFED.

*Tabla 1: LoRAs utilizadas para la generación del conjunto de datos UIBAIFED*

| *LoRA* | Categoría emocional |
|---|---|
| Sad - Facial Expression | Tristeza |
| Look alive! Excited Facial Expression/Emotion | Entusiasmo |
| Ashamed - Embarrassed Facial Expression | Vergüenza |
| Look alive! Smirking Facial Expression/Emotion | *Alegría* |
| Disgusted - Disapproving Facial Expression | Asco |
| Angry - Facial Expression | Ira |
| Shocked - Facial Expression | Sorpresa |
| Pleading Eyes - Facial Expression | Miedo |
| Scared - Facial Expression | Miedo |
| Expressions Helper Realistic | De uso general |

Además, se desarrollaron los *prompts* necesarios para la generación de las micro expresiones que conforman el conjunto de datos. De las 33 micro expresiones descritas por Gary Faigin, solo se logró reproducir un subconjunto de 22, debido a la dificultad de describir ciertas sutilezas mediante modelos generativos.

A continuación, se muestra un ejemplo de los *prompts* generados (positivos y negativos):

*--prompt "White Man, 15y.o, (AngryShouting:0),(angry!!), (((shouting!!!))), <lora:l_ang_ae_sd_64_32:0.9>, Underweight, ((looking at the camera)), hyperrealistic, professional photo, studio lighting, sharp focus, centered on the image, vertical alignment, face, plain grey background"*

*--negative_prompt "((Deformed)), disfigured, hat,(artifacts in eyes, bad iris), ((artifacts in face)), hawaiian clothes, worse quality, low quality, jpeg, pixelated, anime, ((poorly illuminated face)), red eyes, ((bad teeth)), ((body, arms, hands, legs, naked))"*

El *prompt* descrito anteriormente genera la imagen mostrada en la figura 4, que representa a un varón de 15 años de complexión delgada expresando la emoción de ira; concretamente, la micro expresión *Angry Shouting*, según la taxonomía de Faigin.

*Figura 4: Ejemplo de una imagen generada correspondiente a la expresión de ira (varón blanco de 15 años, complexión delgada).*

La estructura de los distintos *prompts* se mantiene de forma consistente, siguiendo el formato que se indica a continuación:

> *"Ethnicity, gender, age, <description of the expression>"*

Dentro de la expresión, se incluye la referencia a la *LoRA* utilizando la siguiente nomenclatura:

> *<lora: (LoRA name):(weight)>*

En esta estructura, el término "*weight*" se refiere a la intensidad de la expresión. Algunas micro expresiones se generan utilizando la misma *LoRA*, pero con descriptores diferentes. Por ejemplo, las microexpresiones *NearlyCrying* y *Sad*, ambas representando la tristeza, se generan con los siguientes dos *prompts* que producen las imágenes mostradas en la figura 5, utilizando la misma *LoRA*.

> *--prompt "Black Woman, 25y.o, (NearlyCrying:0), ((sad mouth)), miserable face, (sad:1.2), <lora:l_sad_se_sd_64_32:1>, Overweight, ((looking at the camera)), hyperrealistic, professional photo, studio lightning, sharp focus, centered on the image, vertical alignment, face, plain grey background"*

> *--prompt "Black Woman, 25y.o, (Sad:0), (sad), (melancholic face), closed lips, small mouth, <lora:l_sad_se_sd_64_32:1>, Overweight, ((looking at the camera)), hyperrealistic, professional photo, studio lightning, sharp focus, centered on the image, vertical alignment, face, plain grey background"*

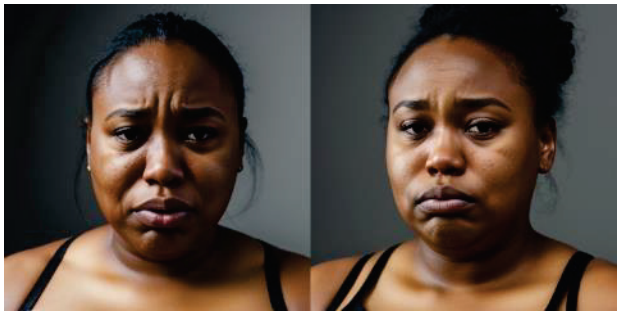


*Figura 5: Expresiones NearlyCrying y Sad generadas con la misma LoRA.*

Los paréntesis y los valores numéricos se utilizan para destacar palabras o frases específicas.

La figura 6 muestra las 22 expresiones generadas para un varón blanco de 15 años. Se desarrolló un *script* automatizado para generar 3960 *prompts*, resultado de la combinación de 2 géneros, 6 etnias, 5 grupos de edad y 3 tipos de complexión corporal, todos organizados según las seis expresiones universales definidas en la clasificación de Ekman (1999).

Las imágenes correspondientes a los *prompts* generados se produjeron utilizando la aplicación *Automatic111* (2022). Debido a la naturaleza aleatoria del proceso de generación de imágenes, no todas las imágenes resultan precisas en el primer intento. Por ello, para cada micro expresión, se

generaron entre 15 y 30 imágenes, con el fin de garantizar la calidad y la coherencia deseadas.

*Figura 6: Ejemplos de las 22 expresiones generadas para un varón blanco de 15 años.*

Las imágenes generadas a partir de los *prompts* especificados fueron seleccionadas manualmente en función de su grado de similitud con las descripciones y representaciones gráficas proporcionadas por Gary Faigin (1990).

La figura 7 ilustra el proceso de comprobación manual para la micro expresión *SlySmile*. La imagen de la izquierda representa la expresión generada a partir del conjunto de datos UIBAIFED, mientras que la imagen de la derecha corresponde a la ilustración de referencia del trabajo de Gary Faigin. Este proceso de selección garantizó que cada imagen representara con precisión la expresión facial prevista y cumpliera con los criterios establecidos.

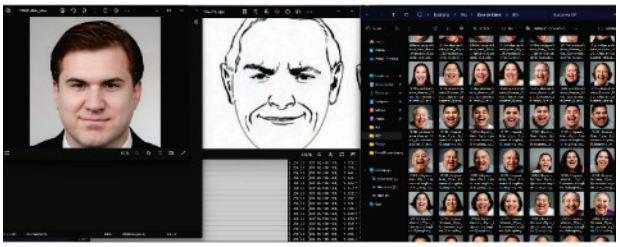


*Figura 7: Proceso de comprobación manual para la expresión SlySmile.*

## 4. El conjunto de datos UIBAIFED

El número total de imágenes que conforman el conjunto de datos es de 2948. Las imágenes generadas a partir de los *prompts* con diferentes tipos de complexión corporal fueron eliminadas debido a las mínimas diferencias observadas entre aquellas etiquetadas como *peso normal* y *bajo peso*.

Se mantuvieron un mayor número de representaciones para las expresiones más complejas, lo que dio lugar a la distribución de imágenes por microexpresión que se muestra en la tabla 2.

*Tabla 2: Distribución de imágenes por microexpresión*

| Expresión Universal | Microexpresión | Número de imágenes |
|---|---|---|
| Neutralidad | Neutral | 123 |
| Ira | AngryShouting | 121 |
| | Mad | 165 |
| | SternnessAnger | 126 |
| Asco | Disdain | 161 |
| | PhysicalRepulsion | 151 |
| Miedo | Afraid | 121 |
| | Terror | 124 |
| | Worried | 135 |
| Alegría | AbashedSmile | 129 |
| | Laughter | 121 |
| | FalseLaughter1 | 120 |
| | FalseLaughter2 | 122 |
| | SmilingOpenMouth | 120 |
| | SmilingClosedMouth | 121 |
| | UproariousLaughter | 119 |
| | SlySmile | 120 |
| Tristeza | CryingClosedMouth | 132 |
| | CryingOpenMouth | 158 |
| | NearlyCrying | 158 |
| | Sad | 155 |
| Sorpresa | Surprise | 146 |

La base de datos está organizada en carpetas, cada una de las cuales contiene imágenes con una resolución de 512×512 píxeles. Existe una carpeta para cada una de las seis expresiones universales según la clasificación de Ekman (1999).

Es importante señalar que la expresión *Desdén* (*Comptent*), que Ekman añadió posteriormente a su clasificación original, está etiquetada en nuestro conjunto de datos como la micro expresión *Disdain*. Esta expresión se incluye dentro de la carpeta *Disgust*, siguiendo el enfoque de clasificación propuesto por Gary Faigin.

Dentro de cada carpeta, las imágenes se nombran siguiendo el siguiente formato:

*Num_ethnicity_gender_age_microexpression.png*

El elemento "Num" representa el número de generación asignado por *Stable Diffusion,* indicando el orden de las imágenes dentro de cada carpeta.

Las imágenes se organizan primero por microexpresión y, posteriormente, por etnia, género y edad.

## 5. Validación del conjunto de datos UIBAIFED

Para validar inicialmente el conjunto de datos UIBAIFED, se empleó un modelo simple de red neuronal convolucional (*CNN, Convolutional Neural Network*) para la clasificación de expresiones faciales.

El modelo recibe como entrada imágenes en escala de grises de 128×128 píxeles, que se procesan a través de tres capas convolucionales. Estas capas van seguidas de una unidad lineal rectificada (*ReLU, Rectified Linear Unit*) y una capa de *max pooling*, utilizadas para extraer las características más relevantes.

La arquitectura también incluye cuatro capas totalmente conectadas que se usan para clasificar las expresiones faciales en una de las 22 micro expresiones objetivo descritas en el conjunto de datos.

La estructura general de la red se muestra en la figura 8.

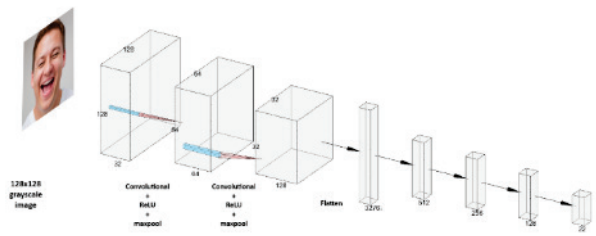



*Figura 8: Estructura del modelo de red neuronal para la validación inicial del conjunto de datos UIBAIFED.*

Para mejorar la capacidad de generalización y evitar el sobreajuste, se aplica una capa *dropout* entre las capas totalmente conectadas. Esta técnica ayuda a que la red aprenda características más robustas, eliminando aleatoriamente algunas unidades durante el entrenamiento, lo que mejora la capacidad del modelo para generalizar a datos no vistos.

El conjunto de datos se divide en subconjuntos de entrenamiento y prueba, utilizando el 67% de los datos para el entrenamiento y el 33% restante para las pruebas.

La distribución de los datos se equilibra únicamente en función de los tipos de microexpresiones, sin considerar otros factores como género, tipo de complexión corporal, etnia o edad en esta etapa de validación. Estos factores serán explorados en estudios futuros.

## 6. Resultados

Tras completar el proceso de entrenamiento, se obtuvo un valor de *loss* cercano a 0.5 y una precisión global (*accuracy*) del 82%. Estos resultados se alcanzaron utilizando el 67% de las imágenes para el entrenamiento, lo que equivale a un total de 1975 imágenes.

El modelo *CNN* entrenado fue evaluado con el conjunto de prueba (compuesto por 1975 imágenes), alcanzando una precisión global del 85.71%.

La matriz de confusión resultante se presenta en la figura 9, mientras que la tabla 3 detalla las métricas de rendimiento obtenidas para cada una de las 22 micro expresiones. La tabla 4 presenta un resumen de las métricas globales de clasificación.

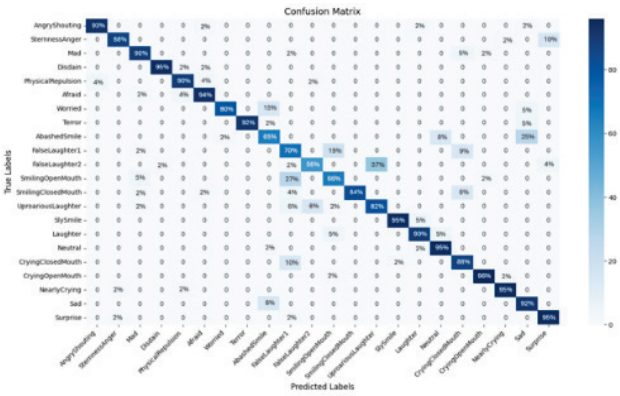



*Figura 9: Matriz de confusión de los resultados de clasificación del modelo CNN.*

*Tabla 3: Métricas de rendimiento para cada microexpresión*

| Microexpresión | Precisión | *Recall* | *F1-Score* | Soporte |
|---|---|---|---|---|
| Neutral | 0.92 | 0.82 | 0.87 | 40 |
| AngryShouting | 0.84 | 0.95 | 0.89 | 43 |
| Mad | 0.77 | 0.82 | 0.80 | 40 |
| SternnessAnger | 0.86 | 0.90 | 0.88 | 40 |
| Disdain | 0.97 | 0.86 | 0.91 | 43 |
| PhysicalRepulsion | 0.88 | 0.87 | 0.87 | 52 |
| Afraid | 0.94 | 0.85 | 0.89 | 53 |
| Terror | 0.93 | 0.97 | 0.95 | 40 |
| Worried | 0.73 | 0.80 | 0.76 | 40 |
| AbashedSmile | 0.68 | 0.65 | 0.67 | 40 |
| Laughter | 0.93 | 0.97 | 0.95 | 40 |
| FalseLaughter1 | 0.57 | 0.72 | 0.63 | 54 |
| FalseLaughter2 | 0.67 | 0.27 | 0.38 | 52 |
| SmilingOpenMouth | 0.61 | 0.61 | 0.61 | 41 |
| SmilingClosedMouth | 1.00 | 0.82 | 0.90 | 50 |
| UproariousLaughter | 0.51 | 0.76 | 0.61 | 51 |
| SlySmile | 0.97 | 0.97 | 0.97 | 40 |
| CryingClosedMouth | 0.72 | 0.69 | 0.71 | 42 |
| CryingOpenMouth | 0.87 | 0.98 | 0.92 | 48 |
| NearlyCrying | 0.97 | 0.93 | 0.95 | 41 |
| Sad | 0.82 | 0.79 | 0.81 | 39 |
| Surprise | 0.85 | 0.80 | 0.82 | 44 |

*Tabla 4: Resumen de las métricas globales de clasificación*

| Métrica | Precisión | *Recall* | *F1-Score* |
|---|---|---|---|
| *accuracy* | | 0.80 | |
| *macro avg* | 0.82 | 0.81 | 0.81 |
| *weighted avg* | 0.81 | 0.80 | 0.80 |

Los resultados de la prueba indican que el modelo *CNN* ha aprendido y generalizado correctamente la mayoría de las expresiones faciales en un conjunto de validación de más de 900 imágenes que no se utilizaron durante el entrenamiento.

La mayoría de las expresiones fáciles alcanzan una precisión superior al 75%. Sin embargo, las expresiones relacionadas con la alegría presentan mayores dificultades de clasificación. En concreto, *AbashedSmile* se clasifica erróneamente en ocasiones como *Sad* o *Worried*, mientras que *FalseLaughter1* se confunde con frecuencia con *CryingOpenMouth*.

Un patrón recurrente observado en todos los ciclos de entrenamiento y prueba es la confusión entre *FalseLaughter2* y *UproariousLaughter*. La principal dificultad para distinguir estas expresiones radica en su gran semejanza visual: ambas representan una boca ampliamente abierta y ojos cerrados o casi cerrados. Este problema ya se había previsto durante el proceso de filtrado de imágenes, donde se observó que las diferencias visuales entre ambas expresiones eran mínimas (véase la figura 10).

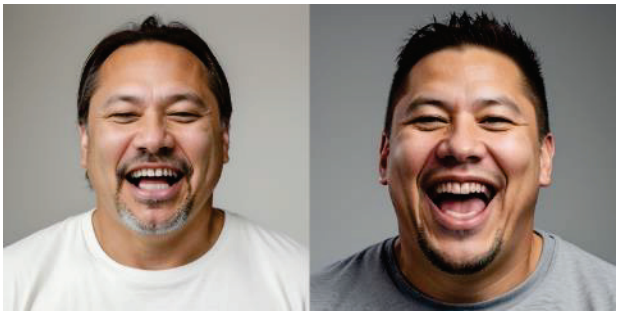


*Figura 10: Comparación entre FalseLaughter2 y UproariousLaugther*

## 7. El experimento con humanos

Una vez comprobado que el conjunto de datos generado puede ser utilizado para entrenar modelos de inteligencia artificial capaces de clasificar imágenes en micro expresiones, se realizó una validación desde la perspectiva del reconocimiento humano para comprobar hasta qué punto el conjunto de datos sintético preserva las características esenciales de las expresiones y verificar que los rasgos generados son reconocibles. La validación consistió en un experimento basado en el reconocimiento, por parte de un grupo de participantes humanos, de un subconjunto de expresiones faciales de UIBAIFED centrado en las imágenes del grupo de edad de 85 años. La elección del subconjunto de datos se basó en el criterio de que, tradicionalmente, el colectivo de personas mayores se encuentra infrarrepresentado en los conjuntos de datos de FER. Además, debido a las modificaciones físicas provocadas por el envejecimiento (como la aparición de arrugas, la flacidez de la piel o la pérdida de firmeza), la externalización de las expresiones faciales en este colectivo puede verse alterada, dificultando, en consecuencia, el reconocimiento de las emociones desde el punto de vista del observador.

### 7.1 Participantes

37 participantes (22 mujeres) fueron reclutados entre el personal del campus de la universidad local y conocidos.
Las edades oscilaron entre 20 y 81 con una media de 52 años (*SD* = 13.69). No hubo requisitos de experiencia previa para participar en el estudio.

### 7.2 El cuestionario

El cuestionario estaba formado por 263 preguntas. Cada pregunta consistía en una imagen de un personaje escogido de forma aleatoria entre las 576 imágenes generadas para el grupo de edad de 85 años del conjunto de datos UIBAIFED. El número final de imágenes correspondientes a cada emoción se mantuvo proporcional al número de imágenes generados por cada emoción en este subconjunto de edad. Junto a la expresión facial se mostraron 7 opciones de respuesta correspondientes a las 6 emociones universales más la neutra: miedo, tristeza, ira, alegría, sorpresa, asco y neutral. Para cada una de las emociones también se proporcionaban algunos de sus sinónimos más habituales (véase la figura 11).

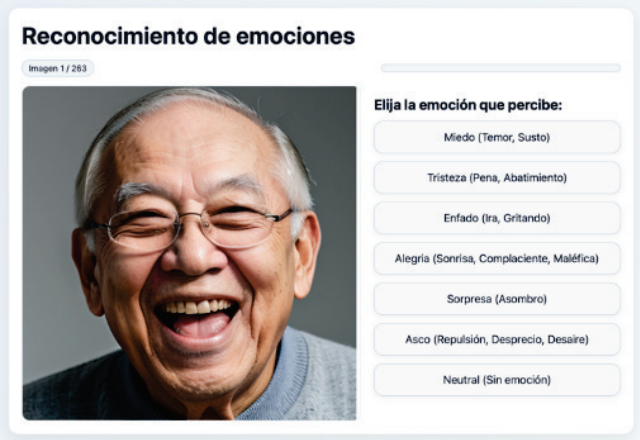



*Figura 11: Ejemplo de una pregunta del cuestionario*

Para cada una de las 263 imágenes se obtuvo una respuesta por participante con el nombre de la emoción percibida. La tabla 5 muestra el número de imágenes por cada emoción utilizadas en el cuestionario. El orden de las preguntas presentadas a los participantes fue aleatorio.

*Tabla 5: Imágenes por emoción utilizadas en el cuestionario*

| Emoción | Número de imágenes |
|---|---|
| *Alegría* | 96 |
| *Tristeza* | 47 |
| *Ira* | 36 |
| *Miedo* | 36 |
| *Asco* | 24 |
| *Sorpresa* | 12 |
| *Neutral* | 12 |

### 7.3 Procedimiento

El cuestionario fue completado online por los participantes. Después de explicar los objetivos del experimento, se instruyó a los participantes para que se sentaran y respondieran el cuestionario con las 263 preguntas. Se les indicó que observaran la imagen presentada en la pregunta e intentaran identificar la emoción correspondiente a la expresión facial mostrada por el personaje. Después de presentar las instrucciones, se proporcionó a los participantes el enlace para poder completar el cuestionario.

Responder al cuestionario tuvo una duración media de unos 18 minutos.

### 7.4 Resultados

En esta sección presentamos los resultados de clasificación obtenidos en términos de precisión (*accuracy*) y matriz de confusión.

La Figura 12 muestra la matriz de confusión obtenida con la clasificación de emociones realizada por los participantes humanos. Los mejores resultados de clasificación fueron para las emociones tristeza (94.6%), neutral (92.8%) y alegría (83.8%), mientras que las peores emociones identificadas fueron las emociones de miedo (33.9%) y asco (34.7%). Las emociones más habitualmente confundidas fueron asco con tristeza (24.9%), miedo con sorpresa (33.5%) y tristeza (23.1%), y sorpresa con miedo (21.4%).

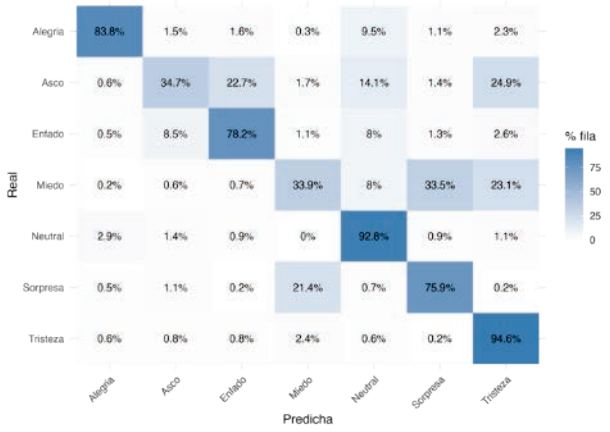



*Figura 12: Matriz de confusión de la clasificación de las emociones realizada por los participantes humanos*

La clasificación de emociones realizada por los participantes obtuvo una precisión global de 0.74. Analizando el efecto del sexo sobre la clasificación de emociones, no se obtuvo significación estadística para el efecto del sexo sobre la precisión ($F_{1,35} = 0.05$, ns).
El análisis de los resultados de clasificación de emociones en función de las micro expresiones, tal como se presenta en la figura 13, permite interpretar con mayor precisión las confusiones observadas en la clasificación.

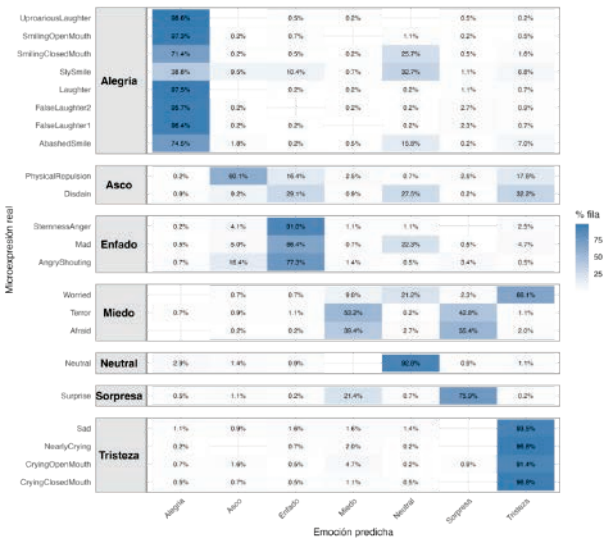



*Figura 13: Matriz de confusión de la clasificación de cada una de micro expresiones en emociones realizada por los participantes humanos*

En el caso de la emoción de alegría, las micro expresiones se caracterizan por la activación del músculo cigomático. En aquellas categorizaciones en las que dicha activación es más intensa (*UproariousLaughter*, *SmilingOpenMouth*, Laughter, *FalseLaughter*, …), la tasa de confusión es mínima. En cambio, las expresiones de baja intensidad, como la risa maliciosa o irónica (*Slysmile*), en las que la elevación de las comisuras labiales es apenas perceptible, muestran una tendencia recurrente a ser clasificadas como expresión neutral.

La emoción de asco incluye dos micro expresiones diferenciadas: la de repulsión física (*PhysicalRepulsion*) y la de desdén (*Disdain*). Los datos indican una confusión clara entre la repulsión y las emociones de enfado y tristeza. Esta confusión puede atribuirse a la relación estrecha entre dichas emociones, así como a la coincidencia en la actividad muscular de la región de las cejas y de la boca, especialmente en expresiones vinculadas al llanto. La expresión de desdén, incorporada más recientemente al conjunto de emociones consideradas universales (Ekman, 1999), presenta una intensidad muscular reducida, lo cual dificulta su reconocimiento en ausencia de un contexto situacional. La confusión con la tristeza puede considerarse, en este sentido, coherente desde un punto de vista perceptivo.

En las expresiones de enfado, la mayor parte de las confusiones se concentra en la micro expresión de menor intensidad (*Mad*), que tiende a ser categorizada como neutral debido a la limitada activación muscular observable.

Las expresiones de miedo constituyen la categoría con menor precisión de clasificación, hecho atribuible principalmente a la falta de contexto situacional. Las confusiones con sorpresa y tristeza resultan coherentes con esta limitación. De manera similar, la confusión entre sorpresa y terror puede explicarse por la similitud en los patrones de activación muscular propios de ambas expresiones. La expresión de preocupación (*Worried*) fue identificada como tristeza por el 66.1% de los participantes, lo que sugiere, no tanto una confusión expresiva como una diferencia en la apreciación subjetiva de la emoción representada.

Por último, las micro expresiones de tristeza alcanzan de media una precisión superior al 94%.

## 8. Conclusiones y trabajo futuro

En este estudio se ha presentado y evaluado UIBAIFED, un nuevo conjunto de datos de expresiones faciales generado por IA compuesto por imágenes en color de alta calidad y realismo, etiquetadas según grupo de edad, género, etnia y expresión facial. El etiquetado sigue las expresiones universales, abarcando un total de 22 micro expresiones basadas en la terminología propuesta por Gary Faigin.

UIBAIFED supone un avance respecto a los conjuntos de datos tradicionales al mejorar la representación étnica, generacional y de complexión física, contribuyendo así a visibilizar la diversidad presente en la población real.

Para validar el conjunto de datos se utilizó una red neuronal convolucional, que alcanzó una precisión del 80% y mostró un rendimiento adecuado en la mayoría de las expresiones faciales. Adicionalmente, se diseñó un experimento con participantes humanos con el fin de evaluar hasta qué punto el conjunto de datos sintético preserva los rasgos distintivos de las expresiones. El experimento se llevó a cabo utilizando un subconjunto de imágenes del grupo de edad de 85 años, un colectivo tradicionalmente marginado en los conjuntos de datos existentes y que, debido a los cambios físicos en la piel provocados por el envejecimiento, puede dificultar el reconocimiento de las expresiones faciales. Los mejores resultados de reconocimiento correspondieron a las emociones de tristeza (94.6%), neutral (92.8%) y alegría (83.8%), mientras que las emociones de miedo (33.9%) y asco (34.7%) presentaron las tasas más bajas, alcanzándose una precisión global del 74%.

El análisis detallado del reconocimiento de micro expresiones permitió identificar las principales fuentes de confusión: la baja intensidad muscular de alguna de las micro expresiones, el solapamiento en los patrones de activación muscular de la región de las cejas y la boca, y la ausencia de un contexto situacional claro.

En comparación con los conjuntos de datos de expresiones faciales existentes, UIBAIFED introduce una contribución fundamental al ofrecer un nivel de etiquetado notablemente más detallado. Hasta donde alcanza nuestro conocimiento, no existe actualmente ninguna base de datos que proporcione un grado similar de granularidad en la anotación de micro expresiones. Además, el uso de la inteligencia artificial generativa facilita la generación de un conjunto de datos más diverso, incluyendo colectivos que tradicionalmente se encuentra infrarrepresentados en los conjuntos de datos de FER.

Este nuevo conjunto de datos abre nuevas oportunidades para la investigación futura, particularmente en el análisis del reconocimiento de expresiones faciales en distintos grupos de edad y etnia. Abordar estos desafíos contribuirá de forma significativa al avance del campo del reconocimiento automático de expresiones faciales.

## Agradecimientos

Este trabajo forma parte del Proyecto PID2022-136779OB-C32 (PLEISAR), financiado por el MICIU/AEI /10.13039/501100011033/ y FEDER, UE, y del Proyecto PID2023-149079OB-I00 (EXPLAINME), financiado por el MICIU/AEI/10.13039/501100011033/ y FEDER, UE.

## Declaración sobre el uso de la IA generativa

Durante la preparación de este trabajo, los autores utilizaron *Stable Diffusion 1.5* y *Realistic Vision V6.0 B1* para la generación de las imágenes que componen el conjunto de datos UIBAIFED.

Asimismo, se empleó *ChatGPT* (versión GPT-4, octubre de 2025) para asistir en la revisión gramatical, ortográfica y de estilo del lenguaje.

Tras el uso de estas herramientas, los autores revisaron y editaron el contenido según lo consideraron necesario y asumen la plena responsabilidad del contenido final de la publicación

## Referencias

Adegun, I. P., & Vadapalli, H. B. (2020). Facial micro-expression recognition: A machine learning approach. *Scientific African, 8*, e00465. https://doi.org/10.1016/j.sciaf.2020.e00465AUTOMATIC1111. (2022). *Stable diffusion web UI* [Software]. https://github.com/AUTOMATIC1111/stable-diffusion-webui

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. En *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (pp. 77–91). PMLR. https://proceedings.mlr.press/v81/buolamwini18a.html

Chen, Y., & Joo, J. (2021). Understanding and mitigating annotation bias in facial expression recognition. *arXiv*. https://doi.org/10.48550/arXiv.2108.08504

Civitai. (s. f.-a). *LoRA Stable Diffusion & Flux AI models*. https://civitai.com/tag/lora

Civitai. (s. f.-b). *Realistic Vision V6.0 B1 – V5.1 Hyper (VAE)*. https://civitai.com/models/4201/realistic-vision-v60-b1

Dominguez-Catena, I., Paternain, D., & Galar, M. (2024). Metrics for dataset demographic bias: A case study on facial expression recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(8), 5209–5226. https://doi.org/10.1109/TPAMI.2024.3361979

Ekman, P. (1999). *Handbook of cognition and emotion*. Wiley.

Faigin, G. (1990). *The artist's complete guide to facial expression*. Watson-Guptill.

Fan, J., Zhou, J., Deng, X., Wang, H., Tao, L., & Kwan, H. K. (2022). Combating uncertainty and class imbalance in facial expression recognition. En *TENCON 2022 – IEEE Region 10 Conference* (pp. 1–4). IEEE. https://doi.org/10.1109/TENCON55691.2022.9977693

Goodfellow, I. J., et al. (2013). Challenges in representation learning: A report on three machine learning contests. *arXiv*. https://arxiv.org/abs/1307.0414

Guerdelli, H., Ferrari, C., Barhoumi, W., Ghazouani, H., & Berretti, S. (2022). Macro- and micro-expressions facial datasets: A survey. *Sensors, 22*(4), 1524. https://doi.org/10.3390/s22041524

Hu, E., et al. (2022). LoRA: Low-rank adaptation of large language models. *arXiv*. https://arxiv.org/abs/2106.09685

Kollias, D., Schulc, A., Hajiyev, E., & Zafeiriou, S. (2020). *Analysing affective behavior in the first ABAW 2020 competition* (p. 643). IEEE. https://doi.org/10.1109/FG47880.2020.00126

Li, S., & Deng, W. (2022). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing, 13*(3), 1195–1215. https://doi.org/10.1109/TAFFC.2020.2981446

Li, S., Deng, W., & Du, J. (2017). Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. En *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2584–2593). IEEE. https://doi.org/10.1109/CVPR.2017.277

Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. En *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (pp. 94–101). IEEE. https://doi.org/10.1109/CVPRW.2010.5543262

Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2019). AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing, 10*(1), 18–31. https://doi.org/10.1109/TAFFC.2017.2740923

Office of Institutional Research. (s. f.). *Race/ethnicity FAQs*. Tufts University. https://provost.tufts.edu/institutionalresearch/race-ethnicity-faq/

Psaroudakis, A., & Kollias, D. (2022). MixAugment & Mixup: Augmentation methods for facial expression recognition. *arXiv*. https://doi.org/10.48550/arXiv.2205.04442

Stable Diffusion AI. (s. f.). *Stable diffusion online – Free AI image generator.* https://stablediffusion.com

Yu, J., Liu, Y., Fan, R., & Sun, G. (2024). MixCut: A data augmentation method for facial expression recognition. *arXiv*. https://doi.org/10.48550/arXiv.2405.10489

Zeng, Y., & Lee, K. (2023). The expressive power of low-rank adaptation. *arXiv*. https://doi.org/10.48550/arXiv.2310.17513

Zhang, Y., Li, Y., Qin, L., Liu, X., & Deng, W. (2023). Leave no stone unturned: Mine extra knowledge for imbalanced facial expression recognition. *arXiv*. https://doi.org/10.48550/arXiv.2310.19636

INTERACCION
Revista Digital de AIPO
Asociación Interacción Persona-Ordenador
Vol. 6, No 2 (2025)
ISSN 2695-6578

# Evaluación efectiva de usabilidad mediante técnicas de análisis y extracción de conocimiento

## Effective usability evaluation by means of analysis and knowledge extraction techniques

**Shuoshuo Li**
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Madrid, España
shuoshuo.li@estudiante.uam.es

**José Antonio Macías Iglesias**
Escuela Politécnica Superior
Universidad Autónoma de Madrid
Madrid, España
j.macias@uam.es

Recibido: 15.11.2025 | Aceptado: 01.12.2025

**Palabras Clave**

Extracción de conocimiento
Análisis de sentimientos
Análisis de audio y video
Pensando en voz alta
Evaluación de usabilidad

**Resumen**

Este trabajo propone la aplicación de técnicas para automatizar evaluaciones de usabilidad basadas en el protocolo *Thinking Aloud* en con el objetivo de superar las limitaciones inherentes a los enfoques manuales tradicionales. Para ello, se realiza una revisión sistemática de la literatura que permitirá identificar avances recientes y vacíos existentes en la aplicación de técnicas automatizadas. El análisis considera tecnologías emergentes como el reconocimiento automático de voz, el procesamiento de lenguaje natural y el análisis multimodal de audio y video, evaluando su potencial para capturar y procesar datos de interacción de manera eficiente y objetiva. Asimismo, se examinan los retos asociados a la integración de estas tecnologías, incluyendo aspectos relacionados con la fiabilidad, la reducción de sesgos y la escalabilidad del proceso. A partir de los hallazgos, se propone el diseño de una herramienta de soporte orientada a la combinación de métodos de aprendizaje y análisis multimodal para optimizar la detección de emociones y la extracción de conocimiento en tiempo real. Esta aproximación busca mejorar la calidad y la eficiencia de las evaluaciones de usabilidad, ofreciendo un marco metodológico que contribuya a la evolución hacia procesos más automatizados y menos dependientes de la intervención humana para aumentar la objetividad.

**Keywords**

Knowledge extraction
Sentiment Analysis
Audio and video processing
Thinking aloud
Usability testing

**Abstract**

This study proposes the automation of the *Thinking Aloud* protocol in usability evaluations, aiming to overcome the limitations of traditional manual approaches. To achieve this, a systematic literature review is conducted to identify recent advances and existing gaps in the application of automated techniques. The analysis examines emerging technologies such as automatic speech recognition, natural language processing, and multimodal audio-video analysis, assessing their potential to capture and process interaction data efficiently and objectively. Furthermore, the review addresses challenges related to the integration of these tools into testing environments, including issues of reliability, bias reduction, and process scalability. Based on these findings, the design of a supporting tool is presented, focusing on the combination of learning methods and multimodal analysis to enhance real-time emotion detection and knowledge extraction. This approach seeks to improve the quality and efficiency of usability evaluations by providing a methodological framework that supports the transition toward more automated processes, reducing human intervention and increasing objectivity.

## 1. Introducción

Las pruebas de usabilidad son un medio fundamental para evaluar la experiencia del usuario y la facilidad de uso de los sistemas interactivos (Hernando & Macías, 2023). El protocolo Thinking Aloud (TA) se considera uno de los métodos más utilizado para realizar investigaciones y evaluaciones de usabilidad (Boren & Ramey, 2000). En particular, este método

se ha adoptado ampliamente debido a su capacidad para registrar de manera intuitiva los procesos de pensamiento y los problemas que enfrentan los usuarios al ejecutar tareas (Boren & Ramey, 2000). Durante una prueba con TA, los usuarios expresan en voz alta sus pensamientos mientras realizan una tarea específica. Este enfoque de "pensar en voz alta mientras se actúa" ayuda a identificar problemas de usabilidad ocultos en el sistema y a comprender los patrones cognitivos de los usuarios (Hertzum & Holmegaard, 2015).

Sin embargo, la mayoría de las pruebas tradicionales con TA se desarrollan manualmente, a través de anotaciones por parte de observadores (Macías & Castells, 2001, 2002; Macías & Culén, 2021). En algunos casos, se utiliza también un proceso de grabación, transcripción y análisis manual, lo que no solo es un proceso que consume mucho tiempo y esfuerzo, sino que también dificulta la eliminación de juicios subjetivos en el análisis (Hertzum & Holmegaard, 2015).

Actualmente, y con el avance de la tecnología, es posible explorar otros enfoques tecnológicos para la predicción de emociones y la extracción automática de características. En los últimos años, se han propuesto y aplicado diversos métodos basados en aprendizaje profundo para la extracción automática de características en la predicción del estado emocional a partir de señales de voz (Jahangir et al., 2021). Sin embargo, algunos estudios han señalado que estos enfoques aún presentan ciertas limitaciones, como una excesiva concentración en el método concurrente, sin considerar en profundidad las diferencias en la aplicación de otros enfoques (McDonald et al., 2012).
Con los avances en el reconocimiento automático de voz (ASR), procesamiento de lenguaje natural (NLP), aprendizaje profundo y análisis de video, la automatización del análisis de grabaciones y videos de usuarios se ha convertido en una dirección clave para poder mejorar la eficiencia y objetividad de las evaluaciones mediante TA. En ese sentido, la investigación actual se centra principalmente en cómo aprovechar las tecnologías existentes para convertir de manera efectiva los datos de voz, texto y video de los usuarios en indicadores cuantificables de emociones y comportamiento, un problema que está recibiendo una creciente atención (Pang & Lee, 2008).

Este artículo tiene como objetivo explorar cómo lograr una evaluación de usabilidad automatizada usando el protocolo TA. A través de la integración de herramientas avanzadas de detección de emociones y extracción de conocimiento, se busca desarrollar un sistema de evaluación eficiente, objetivo y basado en datos. Además, se analizará el estado actual de las tecnologías existentes en el procesamiento de datos de voz, texto y video, así como sus limitaciones (Pang & Lee, 2008). De esta forma, se propone también el diseño de una herramienta de suporte al evaluador, proporcionando un sólido respaldo técnico para la toma de decisiones.

Este artículo se estructura de la siguiente forma. En la Sección 1 introduce el contexto de la investigación. En la Sección 2 se presenta una Revisión Sistemática de la Literatura, formulando las preguntas de investigación y dando contestación a cada una de ellas. En la Sección 3 se discuten los resultados obtenidos, así como las oportunidades de investigación junto con un análisis de las posibles amenazas a la validez identificadas durante el desarrollo del estudio. En la Sección 4 se describe una propuesta de diseño de una herramienta de soporte al evaluador, en base a las indagaciones realizadas en secciones anteriores. Finalmente, en la Sección 5 se presentan las conclusiones del trabajo y las líneas de trabajo futuro que definirán la continuidad de la investigación.

## 2. Análisis de trabajo relacionado

Para investigar sobre las posibilidades de automatización del protocolo TA, se llevará a cabo una SLR (*Systematic Literature Review* o Revisión Sistemática de la Literatura) que permita recopilar y analizar datos utilizando un método repetible y analítico (Kitchenham et al., 2009; Fernández & Macías, 2021). Este enfoque es adecuado para situaciones que requieren un análisis preciso, y es una vía apropiada para identificar y examinar la evidencia existente. La idea es realizar una integración exhaustiva de la literatura relacionada con la automatización de las pruebas de usabilidad basadas en TA, la detección de emociones y la extracción de conocimiento. Esto proporcionará una base teórica sólida para la selección de tecnologías y el diseño de una herramienta de soporte.

### 2.1 Método

En términos generales, la metodología adoptada puede variar entre diferentes disciplinas. El método empleado en este estudio se basa en una versión simplificada de una SLR (Kitchenham & Charters, 2007), con una adaptación específica del marco PICOC (*Población, Intervención, Comparación, Resultados y Contexto*) a nuestro dominio de investigación (Padua, 2010). Esto nos permitirá comprender de manera integral los últimos avances en detección de emociones, extracción de conocimiento y análisis de video dentro de la evaluación automatizada de la usabilidad mediante TA. El método SLR no solo garantiza la repetibilidad y sistematicidad, sino que también permite una recopilación y análisis exhaustivo y riguroso de la literatura existente. A continuación, se presentarán las preguntas de investigación, la cadena de búsqueda, los criterios de inclusión y exclusión, y el cribado y la selección de artículos.

### 2.2 Preguntas de investigación

Se plantean las siguientes preguntas de investigación para guiar la SLR y analizar los resultados obtenidos, en base a los objetivos marcados en la sección anterior:

- RQ1: ¿Qué propuestas existen para el análisis de emociones, principalmente a partir de grabaciones con usuarios, analizando video y audio?

- RQ2: ¿Qué métodos específicos existen para implementar el protocolo Thinking Aloud en estudios de usabilidad, considerando aquellos que puedan ser implementados de manera automática a través de métodos de extracción de conocimiento y de emociones?

- RQ3: ¿Qué tecnologías y herramientas de apoyo existen para la detección de emociones a partir de videos grabados con usuarios?

### 2.3 Selección de palabras clave y fuentes bibliográficas

Para poder encontrar literatura que respalde y responda las preguntas anteriores, se han seleccionado una serie de palabras clave que compondrán la cadena de búsqueda bibliográfica:

Situación experimental y entorno: *Thinking Aloud*
Tratamientos: *AND (Design Thinking OR Usability OR User Experience OR UX OR User-Centered Design OR User-Centred Design)*
Variables de respuesta: *AND (Automatic OR Voice OR Video OR Speech Recognition OR Conversational OR Natural Language Processing OR Intelligent OR Sentiment Analysis OR Transcription)*

Por otro lado, las bases de datos utilizadas para la búsqueda fueron las siguientes:

(1) IEEE Xplore
(2) ACM Digital Library
(3) Scopus
(4) Springer
(5) Google Scholar

### 2.4 Criterios de selección de artículos

Se definieron los siguientes criterios para filtrar los artículos extraídos de las cinco bases de datos mediante la cadena de búsqueda anterior:

- Criterios de inclusión: Los artículos seleccionados deben estar relacionados con evaluaciones de usabilidad que utilicen TA, e incluir procesos de automatización a través de la detección de emociones o la extracción de conocimiento. También, los artículos seleccionados deben estar en inglés o español, y deben haber sido publicados después del 2010 para asegurar que son trabajos recientes.

- Criterios de exclusión: Se descartarán aquellos artículos que no estén relacionados directamente con la temática definida, aquellos que sean antiguos, con contenido en curso, demasiado cortos o poco desarrollados, y aquellos que se encuentren repetidos.

### 2.5 Cribado y selección de artículos

Inicialmente, se obtuvo una alta cantidad de artículos que posteriormente fueron filtrados mediante los criterios de inclusión y exclusión. En la Tabla 1 se muestra un resumen de estos procesamientos iniciales. Para aplicar los criterios de inclusión y exclusión se utilizaron los filtros automáticos que proporcionan las bases de datos. Además, se analizaron los resúmenes, palabras clave y contenido general de los artículos, descartando aquellos que no cumplían con los criterios establecidos. Como podemos ver en la Tabla 1, la cantidad de artículos se redujo significativamente, obteniendo al final un subconjunto más reducido y relevante de artículos.

*Tabla 3: Resultado de la búsqueda de artículos*

| Base de datos | Número inicial de artículos encontrados | Número de artículos tras aplicar los criterios de inclusión y exclusión |
|---|---|---|
| (1) IEEE Xplore | 382 | 25 |
| (2) ACM Digital Library | 1240 | 44 |
| (3) Scopus | 980 | 39 |
| (4) Springer | 1650 | 60 |
| (5) Google Scholar | 32000 | 95 |

Tras otra revisión más concienzuda de los artículos resultantes, finalmente fueron 15 los artículos considerados como primarios, es decir, aquellos relevantes y directamente relacionados con el tema de investigación propuesto, por lo que se utilizarán principalmente para la contestar a las preguntas de investigación planteadas. Además, para ampliar el alcance de la revisión de la literatura, se utilizó el método de la *bola de nieve*. Para ello, tras seleccionar los 15 artículos primarios, se revisaron sus referencias bibliográficas,

encontrando otros 6 artículos adicionales considerados como relevantes para la investigación, lo que aumentó el número de artículos a 21. Estos artículos adicionales enriquecieron la investigación, proporcionando más argumentaciones e información de interés para elaborar este artículo. Por último, se seleccionaron 14 documentos más, a través de búsquedas menos estructuradas en Google. Estos documentos, considerados como *literatura gris*, tratan sobre temas técnicos que son de interés para enriquecer la investigación. Esto permitió finalmente contar con un total de 35 artículos de interés que permitieron abordar cada una de las preguntas de investigación, aportando bibliografía básica para poner en contexto cada una de ellas. Por un lado, para abordar la RQ1 se encontraron artículos que discuten principalmente los métodos de reconocimiento de emociones en video y audio, así como el análisis multimodal de emociones. Para la RQ2 se encontraron artículos que exploran cómo utilizar tecnologías de reconocimiento automático de voz, procesamiento de lenguaje natural y aprendizaje profundo. Por otro lado, para la RQ3 se encontraron artículos que presentan principalmente tecnologías de reconocimiento facial, detección de emociones en tiempo real y herramientas de código abierto relacionadas.

## 3. Resultados y discusión

La revisión en profundidad de la literatura arrojó un resultado alentador para la investigación, y es que, hasta el momento, no se han encontrado soluciones ni propuestas concretas completamente automatizadas para la evaluación de la usabilidad utilizando TA a través de tecnologías de reconocimiento de audio y video para la captura y análisis del pensamiento del usuario. A pesar de los avances en la automatización de algunos aspectos, como el análisis de emociones a partir de audio y video, todavía no existe un sistema integrado que cubra las necesidades del TA de una manera completamente automatizada. Por lo tanto, este campo sigue siendo un área de investigación poco explorada, lo que subraya la importancia de este estudio y su contribución para avanzar hacia una automatización más completa de las pruebas de usabilidad.

A continuación, se analizan cada una de las preguntas de investigación, discutiendo los trabajos encontrados y las soluciones más relevantes aportadas por la literatura.

**(RQ1) ¿Qué propuestas existen para el análisis de emociones, principalmente a partir de grabaciones con usuarios, analizando video y audio?**

Durante mucho tiempo, en la investigación sobre usabilidad ha habido una gran variedad de métodos utilizados para el análisis de emociones basados en grabaciones de usuarios (audio y video). Actualmente, se emplean principalmente los siguientes métodos:

1. Reconocimiento de emociones unimodales en voz. Los investigadores analizan las señales de voz y video de los usuarios, extrayendo características como el volumen, ritmo, energía y tono para detectar emociones. Klaus R. Scherer fue uno de los primeros académicos en investigar la relación entre las características de las señales de voz y la expresión emocional en 1986 (Scherer, 1986). En 2010, Rafael A. Calvo y Sidney D'Mello discutieron las teorías psicológicas de las emociones y revisaron los métodos tradicionales de detección de emociones (como la fisiología, el análisis facial y el análisis de voz), así como los sistemas multimodales emergentes. Calvo y D'Mello (Calvo & D'Mello, 2010) propusieron métodos de análisis de emociones basados en aprendizaje automático y modelos estadísticos. En 2011, El Ayadi y otros concluyeron que las señales de voz son una forma rápida y efectiva de interacción hombre-máquina, enfocándose en el reconocimiento de emociones en la voz (SER) en aplicaciones como la traducción y asistentes inteligentes, y resumieron las técnicas más relevantes, como la extracción de características y los métodos de clasificación (El Ayadi et al., 2011). También se identificaron algunas observaciones sobre los patrones de comportamiento de los usuarios en diferentes condiciones. Por ejemplo, una investigación señala que los analistas emplearon diversos métodos para identificar problemas de usabilidad. En el modo de "trabajo en silencio", los usuarios tendieron a depender más de la búsqueda en el sistema, mientras que en el modo de "pensar en voz alta", este método mostró un carácter más activo y orientado a objetivos (McDonald, 2020). Este hallazgo ofrece una perspectiva valiosa para comprender las características cognitivas del método TA.
2. Métodos de reconocimiento de emociones multimodales. Comenzó en los años 90 y, con el desarrollo de la tecnología, la combinación de voz y video ha mejorado la precisión del reconocimiento de emociones. En 1995, la investigación de la profesora Rosalind Picard sentó las bases para este enfoque (Poria et al., 2017). En 2017, Scherer y Ellgring señalaron que la combinación de voz y video permite un reconocimiento de emociones más efectivo (Scherer & Ellgring, 2007). Por otro

lado, en 2011, Mohammad Soleymani y otros investigaron la base de datos multimodal MAHNOB-HCI, que se utiliza para la investigación de la detección de emociones y etiquetado (Soleymani et al., 2012). Realizaron experimentos utilizando datos de video, movimientos oculares, audio y señales fisiológicas, y demostraron que los métodos multimodales son más precisos que los unimodales. A lo largo del tiempo, los investigadores han comparado diferentes métodos de reconocimiento de emociones. En 2011, Schuller y otros resumieron los avances en el reconocimiento de emociones a partir de la voz (Schuller et al., 2011). En 2017, Scherer y Ellgring estudiaron la expresión multimodal de las emociones y descubrieron que las emociones están influenciadas por el contexto (Scherer & Ellgring, 2007). Estos métodos, combinados con técnicas de aprendizaje profundo y aprendizaje automático, han mejorado la precisión del reconocimiento.

**(RQ2) ¿Qué métodos específicos existen para implementar el protocolo Thinking Aloud en estudios de usabilidad, considerando aquellos que puedan ser implementados de manera automática a través de métodos de extracción de conocimiento y de emociones?**

Para automatizar el TA, se pueden utilizar cinco métodos, cada uno basado en diferentes tecnologías que convierten la voz en texto y luego usan tecnologías avanzadas para realizar el análisis de emociones y la extracción de conocimiento, permitiendo el análisis automático de los informes verbales de los usuarios:

3. Transcripción automática de voz. Utiliza la tecnología de reconocimiento automático de voz para convertir las descripciones verbales en texto. Un estudio preliminar realizado por Nguyen y otros (2017) respalda esta idea, al demostrar que la transcripción de protocolos de pensamiento en voz alta mediante ASR es viable y puede integrarse de manera efectiva en estudios de usabilidad (Kuhn et al., 2024). En 2012, Geoffrey Hinton y otros señalaron que las redes neuronales profundas (DNN) son más eficaces que los métodos tradicionales, mejorando la precisión del reconocimiento de voz, y mencionaron las experiencias de cuatro equipos de investigación exitosos (Hinton et al., 2012).
4. Comprensión del texto y análisis de emociones. Después de la transcripción automática de voz, el texto se introduce en un sistema de procesamiento de lenguaje natural, utilizando modelos de lenguaje pre-entrenados como BERT para el análisis. En 2019, Jacob Devlin y otros señalaron que BERT es un modelo de lenguaje basado en *Transformer*, que aprende representaciones del lenguaje mediante un entrenamiento bidireccional profundo, y es adecuado para diversas tareas NLP (Sun et al., 2020). En comparación con los modelos tradicionales, BERT mostró mejor rendimiento en varias tareas, pudiendo identificar eficazmente las inclinaciones emocionales de los usuarios, apoyando la detección de emociones. Además de estos modelos avanzados, Liu, Li y Wang (2016) propusieron un enfoque basado en procesamiento de lenguaje natural que puede ser usado para el análisis automatizado del protocolo TA, demostrando que es posible identificar patrones problemáticos en los informes verbales de los usuarios de manera automática (Zhang et al., 2024).
5. Extracción de palabras clave y descubrimiento de conocimiento. El uso de tecnologías de aprendizaje profundo permite extraer palabras clave de manera rápida y precisa. En Wang y otros (2019) propusieron un método automático de extracción basado en aprendizaje profundo, que soporta la extracción de palabras clave de las transcripciones, lo que facilita la posterior extracción de conocimiento y el análisis de emociones (Sun et al., 2020). Posteriormente se desarrolló un enfoque específico, que podría ser utilizado para automatizar el TA, orientado a reducir la carga manual y mejorar la precisión del proceso (Li et al., 2025).
6. Modelado de contexto y manejo de textos largos. En 2017, Vaswani y otros concluyeron que *Transformer* puede utilizarse para manejar textos largos y modelar el contexto (Vaswani et al., 2017), lo que permitiría mejorar la comprensión del proceso de pensamiento del usuario y la precisión en la detección de emociones.
7. Explicación del modelo y validación de resultados- En 2016, Ribeiro propuso el modelo LIME, que ayuda a automatizar el proceso de extracción de características emocionales y de conocimiento (Ribeiro et al., 2016). Este método utiliza redes neuronales profundas para la transcripción de voz, BERT para el análisis de emociones, aprendizaje profundo para la extracción de palabras clave, y *Transformer* para el procesamiento del contexto, mejorando así la precisión y la eficiencia de los datos.

Por otro lado, los métodos multimodales pueden ser también de utilidad para en análisis de gestos en el protocolo TA. Si bien la mayor parte del conocimiento se puede extraer a través de la voz, también pueden ser de interés las emociones expresadas por el usuario a través de gestos o expresiones faciales. Sin embargo, la combinación de voz y video ha mejorado la precisión del reconocimiento de emociones, y prueba de ellos son las investigaciones que afirman que la combinación de voz y video permite un reconocimiento de emociones más efectivo (Scherer & Ellgring, 2007), a través de trabajos donde se realizaron experimentos utilizando datos de video, movimientos oculares, audio y señales fisiológicas, demostrando que los métodos multimodales son más precisos que los unimodales (Soleymai et al., 2012). Oros trabajos descubrieron que las emociones están influenciadas por el contexto (Scherer & Ellgring, 2007), y la combinación de técnicas de aprendizaje profundo y aprendizaje automático pueden ser la solución para un reconocimiento automático de emociones más completo en la evaluación automática de la usabilidad mediante TA.

En los últimos años, el desarrollo de las tecnologías de reconocimiento de emociones ha dado un nuevo impulso a las investigaciones en este ámbito. Por ejemplo, algunos estudios han demostrado que los métodos de reconocimiento emocional basados en expresiones faciales pueden alcanzar una tasa de precisión superior al 90 % en condiciones en tiempo real, lo que evidencia su viabilidad y eficacia en aplicaciones prácticas (Abdat et al., 2011). Además, la comunidad académica ha llevado a cabo revisiones sistemáticas sobre la aplicación de la computación afectiva en el reconocimiento de emociones psicológicas, lo cual aporta un sólido respaldo teórico para la comprensión de la percepción emocional en la interacción persona-ordenador (Bakkialakshmi &. Sudalaimuthu, 2021). Otros trabajos han desarrollado programas de reconocimiento de emociones faciales basados en realidad virtual, dirigidos específicamente a la evaluación y tratamiento asistido de personas con esquizofrenia (Souto et al., 2019). Al mismo tiempo, las tecnologías multimodales de reconocimiento emocional sin contacto también están evolucionando rápidamente, y la literatura existente ha explorado sus múltiples aplicaciones, desafíos técnicos, soluciones propuestas y perspectivas futuras (Khan et al., 2024). De la misma forma, en los últimos años, los investigadores también han comenzado a explorar diferentes enfoques tecnológicos, como los métodos de reconocimiento de emociones en el habla basados en aprendizaje profundo (Bhavan et al., 2020), así como el uso de la realidad aumentada para la detección de emociones a partir de expresiones faciales (Bhardwaj, 2023). Estos avances enriquecen aún más el contexto tecnológico en el que se enmarca este artículo.

**(RQ3) ¿Qué tecnologías y herramientas de apoyo existen para la detección de emociones a partir de videos grabados con usuarios?**

Las tecnologías actuales se centran principalmente en el reconocimiento de expresiones faciales en los videos. Para ello, se lleva a cabo la extracción de características faciales de los fotogramas del video para analizar el estado emocional. Las tecnologías y herramientas principales se dividen en los siguientes cuatro enfoques:

8. Métodos de reconocimiento de emociones en video basados en aprendizaje profundo. En 2017, Li y otros propusieron la base de datos RAF-DB y optimizaron el aprendizaje de características mediante el modelo DLP-CNN, demostrando que este modelo supera a los métodos existentes (Li et al., 2022). Ese mismo año, Mollahosseini y otros crearon la base de datos AffectNet, que recopila más de un millón de imágenes faciales, descubriendo que los métodos de Redes Neuronales Profundas (DNN) son más efectivos que los métodos tradicionales. En una investigación previa, los mismos autores habían evidenciado que el uso de arquitecturas DNN más profundas podía mejorar significativamente la precisión en el reconocimiento de expresiones faciales, sentando así una base técnica sólida para sistemas de análisis emocional en video (Mollahosseini et al., 2016).
9. Tecnologías de detección de emociones mediante el reconocimiento en tiempo real de expresiones faciales. En Shuster, M., y otros (2017) investigaron la tecnología de reconocimiento de expresiones faciales en tiempo real. Esta tecnología permite capturar rápidamente los cambios emocionales de los usuarios durante su interacción, asegurando la precisión y la inmediatez de la detección de emociones (Bartlett et al., 2003). En 2019, C. Jiang, Y. Qiu, H. Gao y otros propusieron una plataforma de retroalimentación para usuarios. Utilizando tecnología de aprendizaje profundo y video de alta velocidad, esta plataforma captura y analiza en tiempo real las expresiones faciales de los usuarios para proporcionar retroalimentación emocional. Esto podría ayudar a los diseñadores a optimizar dinámicamente la interfaz de usuario y mejorar los resultados de las pruebas de usabilidad (Jiang et al., 2019). En general, estas tecnologías podrían ser

adecuadas para pruebas de usabilidad en línea y otros escenarios altamente interactivos.

10. Herramientas y plataformas de código abierto. El paquete OpenFace, desarrollado en 2018 por T. Baltrušaitis y otros, es una herramienta de análisis facial de código abierto que puede detectar puntos clave faciales, postura de la cabeza, unidades de acción y seguimiento ocular, y soporta procesamiento en tiempo real. OpenFace tiene una alta precisión y puede ejecutarse en cámaras comunes, siendo aplicable en campos como la interacción persona-ordenador, la computación emocional y el análisis médico (Baltrušaitis et al., 2016).
11. Explicación del modelo y mejoras. En 2017, Selvaraju y otros propusieron el método Grad-CAM, que genera imágenes visuales utilizando información de gradientes, ayudando a entender las áreas clave en el reconocimiento de expresiones faciales (Selvaraju et al., 2017). Las tecnologías actuales, basadas en visión por computadora y aprendizaje profundo, pueden detectar en tiempo real el estado emocional del usuario, proporcionando soporte para las pruebas de usabilidad.

En general, todas las tecnologías y herramientas analizadas parecen prometedoras para su utilización dentro de la interacción con el usuario, con el objetivo de poder detectar información y conocimiento con precisión. Esto permitiría automatizar en análisis de resultados en pruebas de usabilidad basadas en TA. A través de la integración de herramientas avanzadas de detección de emociones y extracción de conocimiento, se podría desarrollar un sistema de evaluación eficiente, objetivo y basado en datos. Esto proporcionará un sólido respaldo técnico para mejorar la experiencia del usuario y el diseño de la interacción (Rojas & Macías, 2015), aportando además una solución automatizada de calidad al proceso (Quintal & Macías, 2021).

### 3.1 Oportunidades de investigación

El análisis realizado a partir de las tres preguntas de investigación revela patrones interesantes de convergencia tecnológica y áreas donde la integración entre diferentes enfoques podría generar avances significativos. Mientras que cada modalidad de análisis (audio, video, texto) ha demostrado capacidades prometedoras de manera individual, la verdadera oportunidad para la automatización completa del protocolo TA reside en la integración inteligente de todos estos enfoques.

La convergencia más significativa se observa en el uso generalizado de arquitecturas de aprendizaje profundo que pueden ser utilizadas en todas las modalidades. Desde los modelos de transformación para procesamiento de texto hasta las redes convolutivas profundas para análisis de video y el procesamiento de audio, existe una base tecnológica común que facilita la integración. Esta convergencia sugiere que es factible desarrollar arquitecturas unificadas que puedan procesar múltiples modalidades de datos de manera coherente y sincronizada, si bien para el caso de TA tiene más sentido el análisis de audio que el análisis de gestos en video.

No obstante, persisten brechas significativas en áreas clave. Primero, la sincronización temporal precisa entre diferentes modalidades sigue siendo un desafío técnico complejo, especialmente cuando se requiere realizar un análisis en tiempo real. Segundo, la interpretación contextual de eventos multimodales en relación con tareas específicas de usabilidad requiere un desarrollo adicional. Tercero, el proceso de adaptación para tener en cuenta las diferencias individuales y culturales respecto a la parte de expresión emocional y de patrones de verbalización presenta desafíos que van más allá de las capacidades técnicas actuales.

#### *3.1.1 Implicaciones para el diseño de herramientas integradas*

Los hallazgos sugieren que el diseño de herramientas para automatizar el protocolo TA debe considerar una arquitectura modular que permita la integración flexible de diferentes componentes tecnológicos. Esta arquitectura debe considerar no solo la precisión técnica de cada componente individual, sino también la coherencia y complementariedad de las percepciones generadas por diferentes modalidades de análisis. Un aspecto particularmente importante es la necesidad de métodos que puedan manejar situaciones donde exista información contradictoria. Por ejemplo, cuando el análisis de audio sugiere frustración, pero el análisis de video indica concentración, la herramienta debe incluir mecanismos para resolver estas inconsistencias de manera inteligente, posiblemente considerando el contexto específico de la tarea y el historial de comportamiento del usuario.

#### *3.1.2 Consideraciones éticas y de privacidad*

La implementación de herramientas de análisis multimodal para la automatización del protocolo TA también plantea importantes consideraciones éticas y de privacidad que deben ser abordadas de manera proactiva. La capacidad de estas herramientas para detectar estados emocionales y cognitivos detallados de los usuarios requiere el desarrollo de marcos éticos claros que protejan la privacidad y

autonomía de los participantes en estudios de usabilidad. Es particularmente importante la consideración de cómo el conocimiento automatizado podría ser utilizado o mal utilizado, especialmente en contextos comerciales donde la información sobre respuestas emocionales de usuarios podría tener valor económico significativo. Las herramientas construidas deben incorporar salvaguardas técnicas y procedimentales que aseguren que los datos emocionales de los usuarios sean utilizados exclusivamente para mejorar la usabilidad y la experiencia del usuario, y no para propósitos de manipulación o explotación comercial.

#### *3.1.3 Desarrollo de estándares y protocolos*

Una necesidad crítica identificada a través de este análisis es el desarrollo de estándares y protocolos específicos para la implementación de herramientas para automatizar el TA. Estos estándares deben abordar aspectos técnicos como la sincronización de datos multimodales, la calibración para diferentes poblaciones de usuarios, y los métodos de validación de resultados automatizados frente a los análisis manuales tradicionales.

Los protocolos también deben especificar mejores prácticas para la recolección de datos, incluyendo configuraciones recomendadas de hardware, procedimientos de configuración experimental, e instrucciones para participantes que optimicen la calidad de los datos capturados sin comprometer la naturalidad de la experiencia del usuario durante las sesiones TA.

#### *3.1.4 Validación y confiabilidad*

Otra área crítica de investigación relacionada es el desarrollo de métodos robustos para validar la confiabilidad y validez de las herramientas de análisis automático en TA. Esto incluye el desarrollo de métricas específicas que puedan comparar de manera significativa los resultados automatizados con los análisis manuales expertos, considerando que diferentes analistas humanos también pueden tener interpretaciones ligeramente diferentes de los mismos datos.
La investigación en validación debe también abordar la cuestión de cuándo y en qué condiciones las herramientas automáticas proporcionan conocimiento que van más allá de lo que puede ser detectado mediante el análisis manual tradicional. Esto podría incluir la capacidad de detectar patrones sutiles en datos multimodales que serían difíciles de percibir por analistas humanos, o la identificación de correlaciones entre diferentes aspectos de la respuesta del usuario que emergen solo cuando se analiza la totalidad de los datos (del *dataset*) de manera sistemática.

#### *3.1.5 Integración con paradigmas de investigación existentes*

Finalmente, se debe abordar cómo las herramientas automáticas podrían ser integradas de manera efectiva en los métodos de investigación de usabilidad existentes. Esto incluye el desarrollo de *workflows* híbridos donde la automatización debe complementar, en lugar de reemplazar completamente, la experiencia del humano, permitiendo que los investigadores se enfoquen en aspectos de alto nivel como la interpretación de resultados y la generación de recomendaciones de diseño.

La integración efectiva también requiere el desarrollo de interfaces de usuario intuitivas que permitan a los investigadores en usabilidad, que pueden no tener experiencia técnica en aprendizaje automático o procesamiento de señales, utilizar e interpretar eficazmente los resultados de las herramientas de automatización. Estas interfaces deben proporcionar tanto resultados de alto nivel como acceso a detalles técnicos cuando sea necesario requerido para una validación o investigación más profunda.

### 3.2 Amenazas a la validez

Con respecto a las amenazas a la validez (Rojas & Macías, 2019) relacionadas con el estudio presentado, el sesgo de publicación, el de selección, junto con la arbitrariedad de la cadena de búsqueda, son las principales amenazas internas a tener en cuenta, mientras que la generalizabilidad puede entenderse como la amenaza externa a considerar en este contexto.

Por un lado, el sesgo de publicación se refiere a la tendencia de que los resultados positivos se publiquen con mayor frecuencia que los negativos, lo que afecta a los resultados de la búsqueda de la literatura. Para reducir este sesgo, se utilizaron múltiples bases de datos (como IEEE Xplore, ACM Digital Library, etc.) y motores de búsqueda (como Google Scholar), además de seleccionar algunos documentos considerados como *literatura gris*, asegurando que se pudiera realizar una búsqueda amplia de investigaciones relacionadas con la automatización del protocolo TA.

Por otro lado, el sesgo de selección se refiere a la inclusión de literatura no relevante debido a criterios de selección demasiado amplios o no estrictos. Para evitar este sesgo, se establecieron criterios claros de inclusión: los artículos deben abordar temas relacionados con la automatización del protocolo TA, la detección de emociones y la extracción de conocimiento, y deben estar escritos en español o inglés, con una fecha de publicación posterior a 2010. Al mismo tiempo, se excluyeron estrictamente aquellos artículos que no se

ajustaran al tema de investigación, que estuvieran repetidos o que tuvieran un contenido demasiado breve, asegurando así que los artículos seleccionados fueran altamente relevantes para el tema de la investigación.

En lo que se refiere a la arbitrariedad de la cadena de búsqueda, ciertamente el diseño de la misma tiene un impacto importante en los resultados obtenidos. Para reducir los problemas de omisión o redundancia causados por combinaciones inadecuadas de palabras clave, estas fueron seleccionadas por los dos autores del artículo, asegurando así que la expresión de búsqueda pudiera capturar con precisión todos los artículos clave relacionados con el tema de investigación planteado.

Finalmente está el problema de la generalizabilidad, es decir, la posibilidad de que los resultados de la investigación no sean aplicables a un campo más amplio. Para reducir este problema se utilizó el método *snowball* o de *la bola de nieve*, aumentando así el número de trabajos relacionados y asegurando que los artículos seleccionados representan mejor el tema de investigación tratado.

Aunque podrían seguir existiendo amenazas a la validez, las acciones llevadas a cabo permiten minimizarlas, proporcionaron una base sólida de literatura para el estudio y análisis de la automatización del protocolo TA a partir de métodos modernos de aprendizaje y extracción automatizada de conocimiento y emociones, que es el objeto de estudio.

## 4. Propuesta de herramienta automatizada

En base a las conclusiones presentadas en la sección anterior, se propone el diseño de una herramienta que permita el análisis automático de evaluaciones de usabilidad a través del protocolo TA, no como método único, sino para la toma de decisiones, de forma que los resultados puedan ser comparados con la percepción de otros observadores humanos. Esto permitiría cubrir las carencias existentes en este ámbito y desarrollar un sistema automatizado que combine datos multimodales y proporcione retroalimentación rápida y eficaz en evaluaciones de usabilidad basadas en TA.

### 4.1 Arquitectura software

La herramienta estará basada principalmente en el análisis de audio; interpretará tanto ficheros de audio como de video que contengan las opiniones de los usuarios grabadas a partir de sesiones TA. La herramienta se implementará mediante una arquitectura modular fundamentada en el patrón Modelo-Vista Controlador. La selección de este patrón arquitectónico responde a la necesidad de integrar flexiblemente diferentes componentes de análisis especializados, mantener la escalabilidad de la herramienta para manejar volúmenes crecientes de datos multimedia, y asegurar la mantenibilidad del código para facilitar futuras extensiones. El patrón MVC se adapta particularmente bien a las características de la herramienta a diseñar debido a la clara separación de responsabilidades que requiere el procesamiento de datos multimedia. El modelo maneja la persistencia de datos complejos incluyendo archivos multimedia, metadatos de análisis, y resultados procesados. La vista gestiona múltiples interfaces especializadas para diferentes tipos de usuarios, desde evaluadores/investigadores individuales hasta administradores de equipos de UX. El controlador coordinará flujos de trabajo complejos que involucran procesamiento asíncrono, integración con APIs externas, y generación de reportes dinámicos.

En la Figura 1 se muestran los principales módulos y etapas funcionales de la herramienta a construir. En concreto, se proponen ocho funcionalidades secuenciales que abarcan desde la entrada que proporciona el evaluador, compuesta por grabaciones de sesiones TA, hasta la salida que proporciona la herramienta, y que estará compuesta por un informe detallado sobre los problemas concretos de usabilidad identificados durante las sesiones TA suministradas como entrada.


*Figura 3: Las ocho etapas funcionales de la herramienta.*

A continuación, se describen estas etapas:

- Entrada y validación: En esta etapa, la herramienta permitirá cargar archivos multimedia, usando formatos estándar de vídeo y audio. También validará dicho formato, la duración y la calidad, haciendo las adaptaciones necesarias para mejorar en lo posible la calidad de la entrada.

- Preprocesamiento: En esta etapa se extraen las pistas de audio, incluyendo las que se encuentren en archivos de video, se normalizará el volumen, y se aplican filtros básicos de ruido.
- Transcripción ASR (*Automatic Speech Recognition*): En esta etapa, el audio procesado en la etapa anterior se convertirá en texto utilizando tecnologías de reconocimiento automático de voz.
- Análisis NLP (*Natural Language Processing*): En esta etapa, el texto obtenido en la etapa anterior se procesa mediante técnicas de procesamiento de lenguaje natural para su *tokenización* y análisis sintáctico.
- Análisis (emocional) de sentimientos: En esta etapa se aplican técnicas de análisis de sentimientos para cuantificar la polaridad emocional de cada segmento obtenido en la etapa anterior.
- Detección de problemas: En esta etapa se aplican algoritmos especializados para detectar patrones específicos que permitan identificar problemas concretos de usabilidad.
- Generación de sugerencias: En esta etapa, los problemas encontrados en la etapa anterior se clasifican por nivel de severidad y se generan las recomendaciones específicas para cada uno de dichos problemas de usabilidad.
- Salida de reporte: En esta etapa, los resultados obtenidos se almacenan en una base de datos, y además se generan visualizaciones y reportes específicos para el usuario final (evaluador), con información sobre el análisis realizado.

Respecto a los detalles técnicos, por un lado, la extracción de audio incluirá un análisis inteligente de los archivos de video para identificar y extraer únicamente las pistas de audio relevantes para el análisis TA. Esto incluye la detección automática de múltiples pistas de audio en archivos complejos, identificación de la pista que contiene verbalizaciones del usuario en contraposición al audio de la propia herramienta o el ruido ambiental, y la preservación de sincronización temporal precisa que permita correlacionar posteriormente transcripciones específicas con momentos exactos en la grabación original.

La normalización de audio implica la utilización de algoritmos sofisticados que ajustan no solamente el volumen general, sino también la dinámica de la señal para optimizar el reconocimiento automático. Esto incluye el uso de compresión dinámica para reducir variaciones extremas de volumen sin eliminar información emocional importante, la ecualización para enfatizar frecuencias de voz humana mientras se suprimen ruidos típicos de grabaciones de campo, y la supresión del ruido que elimina segmentos de silencio o ruido de fondo que pueden interferir con el reconocimiento. La segmentación divide estratégicamente el audio en *chunks* de duración optimizada para su procesamiento a través de APIs de reconocimiento. La duración debe representar un balance cuidadoso entre varios factores, relacionados con la maximización de la precisión de reconocimiento (por ejemplo, los segmentos demasiado cortos pierden contexto, y los que son demasiado largos aumentan errores). Se debe también optimizar el uso de APIs comerciales (que frecuentemente tienen límites de duración por petición), y facilitar el procesamiento paralelo para reducir el tiempo total de análisis.

Por otro lado, el módulo de reconocimiento automático de voz es responsable de la transformación crítica de verbalizaciones auditivas en el texto estructurado que posteriormente puedan ser procesadas por algoritmos especializados de análisis de lenguaje natural para, posteriormente, poder detectar problemas de usabilidad. Esta transformación representa uno de los desafíos técnicos más complejos de la herramienta a construir, ya que debe manejar las características específicas del habla conversacional típica en sesiones del protocolo TA, incluyendo frases falsas, pausas cognitivas, lenguaje informal, y expresiones emocionales espontáneas.

La arquitectura del módulo de reconocimiento automático debe ser diseñada considerando las limitaciones y variabilidades inherentes en las grabaciones de campo típicas de la investigación en UX, donde las condiciones de audio pueden ser subóptimas debido a factores como el ruido ambiental, la calidad variable de los equipos de grabación, los diferentes acentos y patrones de habla de los participantes, y las fluctuaciones en el volumen de voz durante la verbalización de los pensamientos. Esta realidad práctica requiere la implementación de múltiples capas de procesamiento y optimización que van más allá de la simple aplicación de APIs de reconocimiento de voz comerciales. La implementación técnica de este módulo debe integrar múltiples tecnologías complementarias que trabajen en conjunto para optimizar la precisión de la transcripción. Esto es particularmente importante, ya que los diferentes motores ASR tienen fortalezas específicas: algunos son adecuados para reconocimiento de lenguaje conversacional informal, otros para audio con ruido de fondo, y algunos para acentos específicos o terminología técnica. La capacidad de cambiar dinámicamente entre motores según las características detectadas del audio permite optimizar la precisión de transcripción caso por caso.

Otra parte importante es lo relacionado con el análisis de sentimientos y el mapeo a problemas concretos de

usabilidad a detectar. Para ello, se utilizarán herramientas para analizar textos informales y conversacionales, estudiando la polaridad e intensidad tanto positiva como negativa de los sentimientos identificados. Adicionalmente, la herramienta implementará un algoritmo de detección de patrones basado en expresiones regulares optimizadas, utilizado para identificar expresiones lingüísticas específicas relacionadas con problemas conocidos de usabilidad, como se ha indicado en la literatura analizada. Para ello, se realizarán clasificaciones en base a la severidad del problema encontrado y a su factor emocional asociado. La herramienta incluirá también un motor de reglas que permita mapear automáticamente los tipos de problemas detectados a sugerencias específicas de mejora, proporcionando valor inmediato a los usuarios. Durante todo el proceso, se tendrán en cuenta métricas y valores necesarios para la estimación y cálculo de la bondad de las predicciones realizadas.

### 4.2 Prototipo

En la Figura 2 se presenta un prototipo de la herramienta web construida, atendido a las consideraciones narradas en la sección anterior.

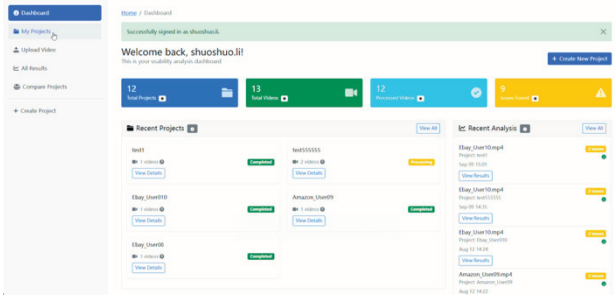



*Figura 2: Prototipo de la herramienta.*

Por un lado, la herramienta permite la gestión de usuarios y proyectos. De esta forma, los evaluadores pueden crear sus propios proyectos de evaluación y subir los videos y audios de las sesiones TA que deseen analizar. Una vez subidos, la herramienta procede a realizar las etapas descritas en la Figura 1, obteniendo como resultado una serie de indicadores y salidas, en forma de *dashboard* o cuadro de mandos, que le servirán al evaluador para la toma de decisiones. Todos los resultados se guardan de manera permanente en una base de datos, de forma que puedan ser accedidos en cualquier momento por parte de los evaluadores. Además, la herramienta permitirá la comparación de distintos tipos de proyectos para identificar problemas de usabilidad comunes a todos ellos.

## 5. Conclusión y trabajo futuro

El protocolo TA es uno de los más utilizados en las pruebas de evaluación de la usabilidad con usuarios, ya que permite, de una forma expresiva, conocer las sensaciones de este mientras interactúa con la aplicación a evaluar, lo que facilita obtener información de primera mano a los evaluadores. La mayoría de las evaluaciones tradicionales mediante TA se desarrollan manualmente, a través de anotaciones que toman los observadores. En algunos casos, se utiliza también un proceso de grabación para su posterior procesamiento, o de transcripción y análisis manual, lo que consume tiempo y esfuerzo.

Con los avances en el reconocimiento automático de voz, procesamiento de lenguaje natural, aprendizaje profundo y análisis de video, la automatización del análisis de grabaciones y videos de usuarios es una solución interesante para mejorar la eficiencia y objetividad de las evaluaciones mediante TA, lo que permitiría un proceso más sistemático de análisis y obtención automática de resultados, logrando pautas para integrar de manera satisfactoria la inteligencia artificial en la interacción con el usuario (Macías 2008; Macías 2012).

Según la literatura existente, no existen todavía propuestas concretas en ese sentido, pero sí hay tecnologías y herramientas que ayudarían a implementar soluciones que permitirían la automatización del protocolo TA y su posterior análisis, especialmente a través del análisis de emociones y de la extracción automática de conocimiento. Esto incluye el reconocimiento automático de voz, que ha permitido mejorar la precisión de la transcripción de voz, haciendo el proceso más eficiente. Por otro lado, el análisis de emociones y procesamiento de lenguaje natural tiene cabida a través de herramientas que permiten analizar el texto transcrito e identificar automáticamente las emociones del usuario. Además, el reconocimiento de emociones multimodales, que permite combinar datos de voz y video, permite capturar con mayor precisión el estado emocional del usuario. Todo ello permitiría extraer y analizar en tiempo real información para asistir en el análisis emocional.

En base a estas indagaciones, se proponen pautas para la construcción de una herramienta de soporte que automatice la detección de problemas de usabilidad en evaluaciones mediante el protocolo TA. La herramienta sigue un flujo lógico compuesto por ocho pasos que permiten cargar información inicial (audio y video) sobre las sesiones TA con usuarios y, como último paso, la emisión de informes de resultados sobre los problemas de usabilidad encontrados, siendo la ejecución del resto de pasos totalmente transparente para el usuario evaluador.

Como líneas futuras para continuar esta investigación, se propone la finalización y testeo de la herramienta de

automatización propuesta, lo que permitiría observar hasta qué punto se puede aumentar la eficiencia, reduciendo al mismo tiempo la intervención humana y mejorando la objetividad. También se pueden plantear otros escenarios que presentan de por sí retos de investigación en esta línea, como la posible falta de precisión en entornos ruidosos y las dificultades técnicas en el análisis emocional en tiempo real cuando se procesan grandes volúmenes de datos. También, la inclusión del análisis facial proveniente de vídeo podría ser explorado como otro canal de entrada de información multimodal, si bien este tipo de información es menos relevante en evaluaciones TA, ya que el audio es la principal fuente de entrada y de conocimiento para evaluaciones basadas en este protocolo.

## Agradecimientos

Esta investigación ha sido subvencionada a través de los proyectos de investigación TED2021-129381B-C21, PID2021-122270OB-I00 y PID2024-155231OB-I00, de la Agencia Estatal de Investigación.

## Referencias

Abdat, F., Maaoui, C., & Pruski, A. (2011). Human–computer interaction using emotion recognition from facial expression. *UKSim 5th European Symposium on Computer Modeling and Simulation* (pp. 196–201). https://doi.org/10.1109/EMS.2011.20

Bakkialakshmi, V. S., & Sudalaimuthu, T. (2021). A survey on affective computing for psychological emotion recognition. *5th International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques (ICEECCOT)* (pp. 480–486). https://doi.org/10.1109/ICEECCOT52851.2021.9707947

Baltrušaitis, T., Robinson, P., & Morency, L. P. (2016). OpenFace: An open source facial behavior analysis toolkit. *IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 1–10). https://doi.org/10.1109/WACV.2016.7477553

Bartlett, M. S., Littlewort, G., Fasel, I., & Movellan, J. R. (2003). Real-time face detection and facial expression recognition: Development and applications to human–computer interaction. *Proceedings of the 2003 Conference on Computer Vision and Pattern Recognition Workshop* (p. 53). https://doi.org/10.1109/CVPRW.2003.10057

Bhavan, A., Sharma, M., Piplani, M., Chauhan, P., Hitkul, S., & Shah, R. R. (2020). Deep learning approaches for speech emotion recognition. In B. Agarwal, R. Nayak, N. Mittal, & S. Patnaik (Eds.), *Deep learning-based approaches for sentiment analysis* (Algorithms for Intelligent Systems). Springer. https://doi.org/10.1007/978-981-15-1216-2_10

Bhardwaj, V., Joshi, A., Bajaj, G., Sharma, V., Rushiya, A., & Bharghavi, S. S. (2023). Emotion detection from facial expressions using augmented reality. *2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA)* (pp. 1–5). https://doi.org/10.1109/ICIRCA57980.2023.10220824

Boren, T., & Ramey, J. (2000). Thinking aloud: Reconciling theory and practice. *IEEE Transactions on Professional Communication, 43*(3), 261–278. https://doi.org/10.1109/47.867942

Calvo, R. A., & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing, 1*(1), 18–37. https://doi.org/10.1109/T-AFFC.2010.7

El Ayadi, M., Kamel, M., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition, 44*(3), 572–587. https://doi.org/10.1016/j.patcog.2010.09.020

Fernández, J., & Macías, J. A. (2021, September). Heuristic-based usability evaluation support: A systematic literature review and comparative study. *In Proceedings of the XXI International Conference on Human-Computer Interaction* (pp. 1-9). https://doi.org/10.1145/3471391.3471395

Hernando, R., & Macías, J. A. (2023). Development of usable applications featuring QR codes for enhancing interaction and acceptance: a case study. *Behaviour & Information Technology*, *42*(4), 360-378. https://doi.org/10.1080/0144929X.2021.2022209

Hertzum, M., & Holmegaard, K. D. (2015). Thinking aloud influences perceived time. *Human Factors, 57*(1), 101–109. https://doi.org/10.1177/0018720814549709

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine, 29*(6), 82–97. https://doi.org/10.1109/MSP.2012.2205597

Jahangir, R., Teh, Y., Wah, H., Faiqa, F., & Mujtaba, G. (2021). Deep learning approaches for speech emotion recognition: State of the art and research challenges. *Multimedia Tools and Applications, 80*(16), 23745–23812. https://doi.org/10.1007/s11042-020-09874-7

Jiang, C., Qiu, Y., Gao, H., Fan, T., Li, K., & Wan, J. (2019). An edge computing platform for intelligent operational monitoring in Internet data centers. *IEEE Access, 7*, 133375–133387. https://doi.org/10.1109/ACCESS.2019.2939614

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering*(EBSE 2007 Technical Report). EBSE. https://www.durham.ac.uk/media/durham-university/departments-/computer-science/research/technical-reports/Guidelines-for-Performing-Systematic-Literature-Reviews-in-Software-Engineering-2007.pdf

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering – A systematic literature review. *Information and Software Technology, 51*(1), 7–15. https://doi.org/10.1016/j.infsof.2008.09.009

Khan, U. A., Xu, Q., Liu, Y., Lagstedt, A., Alamäki, A., & Kauttonen, J. (2024). Exploring contactless techniques in multimodal emotion recognition: Insights into diverse applications, challenges, solutions, and prospects. *Multimedia Systems, 30*(115). https://doi.org/10.1007/s00530-024-01302-2

Kuhn, K., Kersken, V., Reuter, B., Egger, N., & Zimmermann, G. (2024). Measuring the accuracy of automatic speech recognition solutions. *ACM Transactions on Accessible Computing, 16*(3), Article 25, 23 pages. https://doi.org/10.1145/3636513

Li, S., & Deng, W. (2022). Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing, 13*(3), 1195–1215. https://doi.org/10.1109/TAFFC.2020.2981446

Li, S., Huang, X., Wang, T., Zheng, J. & Lajoie, S. (2025). Using text mining and machine learning to predict reasoning activities from think-aloud transcripts in computer-assisted learning. *Journal of Computing in Higher Education, 37*, 477–496. https://doi.org/10.1007/s12528-024-09404-6

Macías, J. A. (2008). Intelligent assistance in authoring dynamically generated web interfaces. *World Wide Web, 11*(2), 253–286. https://doi.org/10.1007/s11280-008-0043-3

Macías, J. A. (2012). Enhancing interaction design on the semantic web: A case study. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42*(6), 1365–1373. https://doi.org/10.1109/TSMCC.2012.2187052

Macías, J. A., & Castells, P. (2001). A generic presentation modeling system for adaptive web-based instructional applications. In *CHI'01 Extended Abstracts on Human Factors in Computing Systems* (pp. 349-350). https://doi.org/10.1145/634067.63427

Macías, J. A., & Castells, P. (2002). Tailoring dynamic ontology-driven web documents by demonstration. In *Proceedings Sixth International Conference on Information Visualisation* (pp. 535-540). IEEE. https://doi.org/10.1109/IV.2002.1028826

Macías, J. A., & Culén, A. L. (2021). Enhancing decision-making in user-centered web development: a methodology for card-sorting analysis. *World Wide Web, 24*(6), 2099-2137. https://doi.org/10.1007/s11280-021-00950-y

McDonald, S., Cockton, G., & Irons, A. (2020). The impact of thinking-aloud on usability inspection. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW1), 1–22. https://doi.org/10.1145/3397876

McDonald, S., Edwards, H. M., & Zhao, T. (2012). Exploring think-alouds in usability testing: An international survey. *IEEE Transactions on Professional Communication, 55*(2), 2–19. https://doi.org/10.1109/TPC.2011.2182569

Mollahosseini, A., Hasani, B., & Mahoor, M. H. (2016). Going deeper in facial expression recognition using deep neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2016)* (pp. 1–6). https://doi.org/10.1109/CVPRW.2016.7477553

Padua, A. G. (2010). Propuesta de un proceso de revisión sistemática de experimentos en ingeniería del software. *Proceedings of the 13th Ibero-American Conference on Software Engineering (CIbSE 2010)* (pp. 313–318). Universidad Politécnica de Madrid.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1–2), 1–135. https://doi.org/10.1561/1500000011

Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affect analysis: From unimodal analysis to multimodal fusion. *Information Fusion, 37*, 98–125. https://doi.org/10.1016/j.inffus.2017.02.003

Quintal, C., & Macías, J. A. (2021). Measuring and improving the quality of development processes based on usability and accessibility. *Universal Access in the Information Society, 20*(2), 203–221. https://doi.org/10.1007/s10209-020-00726-7

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*(pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

Rojas, L. A., & Macías, J. A. (2015). An agile information-architecture-driven approach for the development of user-centered interactive software. *Proceedings of the XVI International Conference on Human–Computer Interaction* (Article No. 50, pp. 1–8). https://doi.org/10.1145/2829875.2829919

Rojas, L. A., & Macías, J. A. (2019). Toward collisions produced in requirements rankings: A qualitative approach and experimental study. *Journal of Systems and Software, 158*, 110417. https://doi.org/10.1016/j.jss.2019.110417

Scherer, K. R. (1986). Vocal communication of emotion: A review of research paradigms. *Speech Communication, 5*(1–2), 1–49. https://doi.org/10.1016/0167-6393(86)90070-X

Scherer, K. R., & Ellgring, H. (2007). Multimodal expression of emotion: Affect programs or componential appraisal patterns? *Emotion, 7*(1), 158–171. https://doi.org/10.1037/1528-3542.7.1.158

Schuller, B., Batliner, A., Steidl, S., & Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication, 53*(9–10), 1062–1087. https://doi.org/10.1016/j.specom.2011.01.011

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *2017 IEEE International Conference on Computer Vision (ICCV)*(pp. 618–626). https://doi.org/10.1109/ICCV.2017.74

Soleymani, M., Lichtenauer, J., Pun, T., & Pantic, M. (2012). A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing, 3*(1), 42–55. https://doi.org/10.1109/T-AFFC.2011.25

Souto, T., Silva, H., Leite, A., Baptista, A., Queirós, C., & Marques, A. (2019). Facial emotion recognition: Virtual reality program for facial emotion recognition—a trial program targeted at individuals with schizophrenia. *Rehabilitation Counseling Bulletin, 63*(2), 79–90. https://doi.org/10.1177/0034355219847284

Sun, Y., Qiu, H., Zheng, Y., Wang, Z., & Zhang, C. (2020). SIFRank: A new baseline for unsupervised keyphrase extraction based on pre-trained language model. *IEEE Access, 8*, 10896–10906. https://doi.org/10.1109/ACCESS.2020.2965087

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008.

Zhang, J., Borchers, C., Aleven, V., & Baker, R. S. (2024). Using large language models to detect self-regulated learning in think-aloud protocols. *Proceedings of the 17th International Conference on Educational Data Mining* (pp. 157–168). International Educational Data Mining Society. https://doi.org/10.5281/zenodo.12729790

INTERACCION
Revista Digital de AIPO
Asociación Interacción Persona-Ordenador
Vol. 6, No 2 (2025)
ISSN 2695-6578

# Exploración de las preferencias de usuarios expertos sobre las regiones de importancia como explicaciones en clasificación de actividades en vídeo

## Exploring expert user preferences regarding importance heatmaps as explanations in video activity classification

**F. Xavier Gaya-Morey**
Ciencias Matemáticas e Informática
Universitat de les Illes Balears
Palma, Islas Baleares, España
francesc-xavier.gaya@uib.es

**Jose M. Buades-Rubio**
Ciencias Matemáticas e Informática
Universitat de les Illes Balears
Palma, Islas Baleares, España
josemaria.buades@uib.es

**Scott MacKenzie**
Electrical Engineering and Computer Science
York University
Toronto, Ontario, Canadá
mack@yorku.ca

**Raquel Lacuesta**
Informática e Ingeniería de Sistemas
Universidad de Zaragoza I3A
Teruel, Aragón, España
lacuesta@unizar.es

**Cristina Manresa-Yee**
Ciencias Matemáticas e Informática
Universitat de les Illes Balears
Palma, Islas Baleares, España
cristina.manresa@uib.es

Recibido: 16.11.2025 | Aceptado: 01.12.2025

**Palabras Clave**

inteligencia artificial explicable
evaluación
XAI centrado en el ser humano
métodos XAI basados en vídeo

**Resumen**

Aunque existen numerosos métodos de inteligencia artificial explicable (XAI), todavía hay una falta de estudios que analicen cómo los usuarios perciben la explicabilidad y la confiabilidad que estos ofrecen. Consecuentemente, es difícil determinar cuáles son los métodos de XAI más adecuados en función de las preferencias y necesidades de los usuarios. En este trabajo, usuarios expertos en IA evaluaron seis métodos XAI basados en perturbación, aplicados a través de tres redes y dos conjuntos de datos para el reconocimiento de actividades en vídeo. Para ello, se pidió a los expertos puntuar cómo de razonables fueron las explicaciones, en base a las regiones del vídeo señaladas como importantes. Los resultados muestran la preferencia por el método RISE adaptado a vídeo, mientras que identifican el método de predictores univariados adaptado a vídeo como el menos razonable. Estos hallazgos ofrecen a investigadores y profesionales una visión sobre los métodos de XAI preferidos en vídeo, al tiempo que amplían la comprensión de la explicabilidad de la IA desde una perspectiva centrada en el usuario.

**Keywords**

explainable artificial intelligence
evaluation
human-centered XAI
video-based XAI methods

**Abstract**

Many explainable artificial intelligence (XAI) methods exist; however, there is a lack of user evaluations on explainability or trustworthiness. Consequently, it remains unclear which XAI methods are appropriate based on users and their preferences. In this study, AI experts evaluated six removal-based XAI methods applied across three networks and two datasets for video-based activity recognition. For this purpose, the experts scored the reasonableness of the explanations, based on the video regions indicated as the most relevant. Experts consistently preferred the video-adapted RISE method, while identifying the video-adapted univariate predictors method as the least preferred. These findings provide insight for researchers and practitioners on the preferred XAI methods to use with videos, while also expanding the understanding of XAI methods from a human perspective.

## 1. Introducción

La creciente presencia de la inteligencia artificial (IA) en múltiples dominios (Dong et al., 2021), pone de relieve la importancia de garantizar su explicabilidad, de modo que las decisiones automatizadas puedan entenderse, evaluarse y, cuando proceda, cuestionarse por los usuarios. Tal como dicta el marco de trabajo HCAI (Human-Centered Artificial Intelligence framework), los métodos que identifican cuándo es necesaria la acción automática y cuándo la humana, y que evitan el exceso de peso de estas acciones son más propensas a producir diseños fiables, seguros y confiables (Shneiderman, 2020).

Desde la llegada del aprendizaje profundo (d*eep learning)*, se han desarrollado numerosos métodos de IA explicable (XAI) con el propósito de hacer inteligible la toma de decisiones de los modelos a los usuarios humanos (Adadi & Berrada, 2018; Barredo Arrieta et al., 2020). Sin embargo, pese al notable avance técnico en técnicas de explicabilidad, todavía es limitado el número de estudios que investigan cómo los usuarios finales perciben estas explicaciones, qué necesidades de explicabilidad tienen y en qué contextos las valoran positivamente.

En este sentido, resulta imprescindible no sólo diseñar nuevos métodos de XAI, sino también incorporar al usuario final desde el inicio del proceso de diseño de las explicaciones. De hecho, la literatura ha comenzado a adoptar un enfoque de IA explicable centrada en el ser humano (Human-Centred XAI, HCXAI), que integra factores humanos en la investigación y desarrollo de explicaciones de IA (Schoonderwoerd et al., 2021; Hong y Park, 2025; Ridley, 2025).

En consecuencia, avanzar en el desarrollo de HCXAI implica alinear las explicaciones de IA con los usuarios específicos, considerando sus niveles de especialización, sus tareas o los entornos de uso, y adaptar las presentaciones al contexto. Esta orientación promueve no sólo la comprensibilidad y la usabilidad, sino también la confianza, la colaboración humano-IA y, la adopción ética y eficaz de los sistemas inteligentes. Estudios recientes muestran además que los criterios para una explicación significativa no se limitan a la fidelidad técnica del modelo, sino que incluyen dimensiones como ser comprensible, accionable, concisa, coherente con la tarea del usuario y adaptada a sus expectativas (Kim et al., 2024).

Aunque las ciencias sociales han estudiado ampliamente los procesos mediante los cuales las personas generan y comprenden explicaciones, la investigación en XAI suele fundamentarse en la intuición de los propios investigadores acerca de qué constituye una "buena" explicación (Miller, 2019). Esta estrategia tiende a pasar por alto aspectos esenciales como la comprensión humana, los perfiles y necesidades de los destinatarios y los factores contextuales en torno a la explicación. Estudios recientes evidencian una baja utilización de los métodos centrados en el ser humano en el diseño de sistemas XAI (Kaplan et al., 2024; Mohseni et al., 2018; Rong et al., 2024). La literatura indica que este tipo de enfoques resultan útiles para orientar las decisiones técnicas impulsadas por las necesidades y perspectivas de los usuarios, al tiempo que permiten identificar limitaciones en los métodos existentes y proporcionar marcos conceptuales que promuevan una XAI compatible con los humanos (Liao & Varshney, 2022).

Además, a nivel de evaluación de XAI, existe una falta de marcos teóricos, metodológicos y métricas estandarizadas que permitan evaluar en qué medida los métodos de XAI ofrecen una explicabilidad útil para las personas (Floridi et al., 2018; Hoffman et al., 2019; Miró-Nicolau et al., 2024). Las escasas evaluaciones empíricas con usuarios disponibles suelen carecer de fundamentos provenientes de las ciencias cognitivas y sociales (Rong et al., 2024) y no siguen protocolos sistemáticos para medir, cuantificar y comparar la explicabilidad de los sistemas de IA (Burkart & Huber, 2021). Además, los estudios que evalúan empíricamente los métodos XAI con tareas, usuarios y contextos específicos muestran diferentes necesidades y preferencias de los usuarios (Dodge et al., 2019; Ehsan et al., 2024; Szymanski et al., 2021). En esta misma dirección, (Wells & Bednarz, 2021) llevaron a cabo una revisión sistemática examinando estudios XAI con un especial interés en el usuario, revelando que muchos estudios no involucraban a usuarios, e incluso cuando se realizaban pruebas con usuarios, a menudo se omitían detalles clave, como el número de participantes, los métodos de reclutamiento o el nivel de experiencia de los participantes en aprendizaje automático. Esto limita la transparencia y la reproducibilidad de las evaluaciones.

AL trabajar específicamente con datos visuales (es decir, imágenes o vídeos), existen evaluaciones de métodos XAI con usuarios para imágenes (Aechtner et al., 2022; Alqaraawi et al., 2020; Heimerl et al., 2020; Manresa-Yee et al., 2024), pero, hasta donde sabemos, no hay trabajos que aborden vídeos. Por lo tanto, es necesario investigar para comprender completamente el impacto de los métodos XAI para vídeo y la efectividad de las explicaciones.

El objetivo de este trabajo es realizar un estudio cuantitativo con expertos que evalúa seis métodos XAI basados en perturbación en vídeo, aplicados sobre tres redes y dos

conjuntos de datos. Para lograr esto, se adaptan seis métodos XAI, ampliamente utilizados originalmente para explicaciones locales basadas en imágenes, al dominio del vídeo. Para entender las diferencias, se generan explicaciones para tres redes con arquitecturas variadas, incluyendo *transformers* y modelos convolucionales. Además, se utilizan dos conjuntos de datos para el reconocimiento de acciones humanas disponibles públicamente: uno grabado en un entorno controlado y otro que comprende vídeos de entornos no controlados extraídos de YouTube.

Tras cuantificar las preferencias de los usuarios respecto a los seis métodos XAI, los resultados muestran acuerdo tanto en los métodos preferidos como con los menos preferidos. Estos hallazgos proporcionan, a futuros investigadores y profesionales, guías de diseño concretas, respaldadas por las elecciones de los usuarios.

El artículo se organiza de la siguiente manera: la Sección 2 proporciona una revisión de los conceptos clave relacionados. La Sección 3 detalla el sistema impulsado por IA, incluidos los conjuntos de datos, las redes neuronales y los métodos XAI utilizados. La Sección 4 describe la metodología, cubriendo los participantes, el aparato, el procedimiento y el diseño del estudio. Los resultados se presentan en la Sección 5, seguidos de una discusión en la Sección 6. Finalmente, la Sección 8 concluye y destaca posibles direcciones para futuras investigaciones.

## 2. Trabajo relacionado

### 2.1. Inteligencia artificial explicable centrada en el ser humano

XAI comprende un conjunto de métodos y técnicas orientados a mejorar la transparencia y la interpretabilidad de las decisiones y del interno de los modelos de inteligencia artificial. A medida que estos sistemas se vuelven más complejos, especialmente en el ámbito del aprendizaje profundo, comprender cómo producen sus resultados se vuelve más complejo. Las técnicas de XAI buscan reducir esta dificultad al ofrecer información sobre el comportamiento del modelo, lo que facilita que los usuarios interpreten, evalúen y confíen en las decisiones generadas por la IA (Adadi & Berrada, 2018; Barredo Arrieta et al., 2020).

Las técnicas de explicabilidad se categorizan generalmente a lo largo de varias dimensiones clave: enfoque en datos o en modelo, explicaciones directas o *post-hoc*, alcance global o local, y presentación estática o interactiva (Arya et al., 2020). Primero, las explicaciones pueden tener como objetivo aclarar las propiedades de los datos de entrada o el comportamiento del modelo en sí. Al explicar el modelo, la distinción es entre modelos directamente interpretables (por ejemplo, regresión lineal, o árboles de decisión) y la explicabilidad *post-hoc*, que se aplica después de que se entrene el modelo. Además, las explicaciones pueden dirigirse a predicciones individuales (local) o al comportamiento del modelo en su conjunto (global). Finalmente, las explicaciones pueden ser estáticas o, como recomienda Miller (2019), diseñadas para apoyar la participación interactiva del usuario para una comprensión más profunda.

La XAI centrada en el ser humano se basa no solo en la explicabilidad técnica, sino también en alinear las explicaciones con las necesidades humanas, los procesos cognitivos y el contexto (Barda et al., 2020; Liao et al., 2020). En lugar de asumir que una explicación es suficiente, la XAI centrada en el ser humano enfatiza la usabilidad, la interpretabilidad y la relevancia para diversos usuarios, incluidos los no expertos (Lopes et al., 2022). El objetivo es crear explicaciones que sean intuitivas y conscientes del contexto y, por lo tanto, apoyen la toma de decisiones para mejorar la colaboración entre humanos y sistemas de IA (Ehsan et al., 2022; Liao & Varshney, 2022). Esta perspectiva reconoce que la efectividad de una explicación depende tanto del usuario como del método en sí.

### 2.2. XAI aplicado a datos de vídeo

Mientras que los métodos XAI basados en imágenes se han estudiado exhaustivamente en la literatura (por ejemplo, Lundberg & Lee, 2017; Petsiuk et al., 2018; Ribeiro et al., 2016), los métodos XAI basados en vídeo, particularmente los agnósticos al modelo, permanecen relativamente inexplorados, en parte debido a los desafíos únicos que plantea el vídeo, como la mayor dimensionalidad. Sin embargo, los métodos agnósticos del modelo son valiosos porque ofrecen flexibilidad y amplia aplicabilidad para escenarios del mundo real.

Los enfoques agnósticos más empleados se basan en la eliminación o modificación de la entrada (*removal-based explanations*). Su objetivo común es estimar la relevancia de las características de entrada analizando los cambios en las predicciones cuando se perturban o eliminan partes de la entrada. Dado que las modificaciones se realizan directamente sobre el espacio de entrada, estos métodos no dependen del modelo, y la importancia de cada característica se infiere únicamente a partir de la variación en la salida. En el caso de los datos de vídeo, la elevada dimensionalidad y la naturaleza espacio-temporal del contenido obligan a introducir adaptaciones, como el uso de grupos de píxeles o

regiones (superpíxeles) y la extensión del análisis a lo largo del tiempo.

Entre los métodos agnósticos más representativos (y los empleados en nuestros experimentos) se encuentran LIME, SHAP, RISE, LOCO, *univariate predictors* y *occlusion sensitivity*.

LIME (*Local Interpretable Model-agnostic Explanations*) explica las predicciones de un clasificador mediante un modelo interpretable local, como una regresión lineal o un árbol de decisión. Particiona la imagen (o fotograma) en regiones llamadas superpíxeles y genera muestras perturbadas ocluyendo aleatoriamente parte de ellas. La relevancia de cada región se estima en función del cambio observado en la confianza de la predicción (Ribeiro et al., 2016).

SHAP (*SHapley Additive exPlanations*) introduce los valores de Shapley como medida unificada de importancia de las características. Estos valores representan el cambio esperado en la predicción del modelo al considerar o excluir una característica. Kernel SHAP, una variante eficiente y aproximada, combina elementos de LIME para garantizar exactitud local y coherencia en la asignación de relevancias (Lundberg y Lee, 2017).

RISE (*Randomized Input Sampling for Explanation*) genera máscaras binarias aleatorias sobre una cuadrícula de baja resolución, que posteriormente se interpola hasta las dimensiones del fotograma. Cada muestra se obtiene ocluyendo diferentes regiones y midiendo el efecto en la probabilidad de clase. La relevancia final se calcula promediando las predicciones en aquellas muestras donde las regiones permanecen visibles, produciendo un mapa de calor de relevancia positiva (Petsiuk et al., 2018).

LOCO (*Leave-One-Covariate-Out*) evalúa la influencia de cada característica eliminándola del conjunto de entrenamiento y observando la variación resultante en las predicciones. A diferencia de los anteriores, proporciona explicaciones globales del modelo en lugar de locales (Lei et al., 2018).

Los *univariate predictors* proponen evaluar el impacto de cada variable de manera independiente, optimizando la interpretabilidad y reduciendo la complejidad computacional. Como LOCO, este enfoque ofrece explicaciones globales (Guyon & Elisseeff, 2003).

*Occlusion sensitivity* calcula la importancia de los píxeles desplazando un parche (por ejemplo, gris) sobre la imagen y midiendo el cambio en la confianza de la predicción. En el contexto del vídeo, este método se extiende añadiendo una dimensión temporal al kernel usado para la oclusión, lo que permite capturar la relevancia espaciotemporal y adaptarse mejor a secuencias con cortes, movimientos de cámara u objetos que entran o salen del campo de visión (Zeiler & Fergus, 2014).

Estos métodos, con las adaptaciones pertinentes para ser extrapolados a su uso en vídeo (Gaya-Morey et al., 2024) constituyen las principales aproximaciones agnósticas utilizadas actualmente para proporcionar explicaciones interpretables en tareas de reconocimiento basadas en vídeo.

### 2.3. Estudios de usuario en datos visuales

La mayoría del trabajo existente en evaluación de métodos XAI se centra en métricas automáticas, a menudo pasando por alto cómo los usuarios reales interpretan, confían o se benefician de estas explicaciones (Miller, 2019). En consecuencia, relativamente pocos estudios de usuario evalúan explicaciones de imágenes y vídeos con participantes humanos. Esta brecha es especialmente pronunciada en el dominio del vídeo, donde la dimensión temporal añade complejidad a la interpretación humana. Evaluar las explicaciones con usuarios reales es crucial para comprender su utilidad práctica, mejorar las opciones de diseño y garantizar que dichos sistemas se alineen con el razonamiento y la toma de decisiones humanos.

Con respecto a los estudios de usuario sobre métodos XAI aplicados a imágenes, Aechtner et al. (2022) estudiaron la percepción de los usuarios sobre las explicaciones locales frente a las globales, mostrando la preferencia por las explicaciones locales de los usuarios poco habituados al uso de IA. Manresa-Yee et al. (2024) también estudiaron las explicaciones locales frente a las globales, involucrando a 104 usuarios en un estudio que analizaba aspectos como la confianza percibida o la comprensión. Se observaron puntuaciones más altas para las combinaciones de ambas explicaciones. Alqaraawi et al. (2020) investigaron el rendimiento de los mapas de saliencia en un estudio con usuarios, mostrando una preferencia por LRP, y señalaron la ayuda limitada de las explicaciones para predecir la salida de la red para nuevas imágenes o para identificar las características de la imagen a las que el sistema es sensible.

Selvaraju et al. (2017) exploraron si las explicaciones de Grad-CAM ayudaron a los usuarios a establecer una confianza adecuada en las predicciones. Sus resultados mostraron que Grad-CAM permitió a los usuarios no entrenados diferenciar con éxito una red profunda "más fuerte" de una "más débil", incluso cuando producían predicciones idénticas.

En nuestra revisión de la literatura para XAI aplicado a vídeo, no se encontró ningún trabajo que evalúe o compare diferentes explicaciones en vídeos desde una perspectiva humana.

## 3. Sistema de reconocimiento de actividades y métodos XAI

Para evaluar las preferencias del usuario por los métodos XAI de vídeo, se creó un conjunto de que combina tres redes, dos conjuntos de datos y seis métodos XAI. Esto permitió introducir variaciones en las explicaciones, para identificar la influencia de estos tres factores.

### 3.1. Conjuntos de datos

Se seleccionaron dos conjuntos de datos con características distintas para entrenar los modelos y evaluar los métodos XAI: Kinetics 400 (Kay et al., 2017) y EtriActivity3D (Jang et al., 2020).

El conjunto de datos Kinetics 400 es una colección a gran escala de vídeos de YouTube que cubre 400 categorías, con al menos 400 videoclips por clase. El conjunto de datos se centra en diversas actividades humanas, incluidas las interacciones entre personas y las interacciones con objetos. Presenta una amplia variedad de participantes, entornos y objetos, junto con desafíos como movimientos de cámara y cortes dentro del mismo clip, lo que contribuye a su complejidad.

En contraste, EtriActivity3D es un conjunto de datos más especializado que contiene 112.620 vídeos, clasificados en 55 actividades. Se centra en tareas cotidianas realizadas por 100 individuos, la mitad de los cuales tienen más de 64 años, lo que permite el estudio sobre demografía de edad avanzada. Los vídeos se capturaron en entornos domésticos, incluyendo múltiples habitaciones, y desde ocho cámaras fijas, asegurando una grabación estable y sin cortes para cada clip. Se trata, por tanto, de una configuración muy controlada que permite obtener un conjunto de datos menos complejo que Kinetics 400.

### 3.2. Redes neuronales

Se utilizaron tres redes: TimeSformer (Bertasius et al., 2021), TANet (Liu et al., 2021) y TPN (Yang et al., 2020). Estas redes son representantes de diferentes arquitecturas neuronales, como los *transformers* y las redes convolucionales, lo que nos permite explorar si dicha arquitectura tiene un impacto en las preferencias de los usuarios. La elección de las redes se justifica por su rendimiento en tareas de clasificación de acciones y su disponibilidad pública dentro de la caja de herramientas de código abierto basada en PyTorch MMAction2 para el análisis de vídeo (OpenMMLab, 2020).

TimeSformer, una variante del *Vision Transformer*, captura características espaciotemporales procesando parches a nivel de fotograma. TANet incorpora un Módulo Adaptativo Temporal (TAM) dentro de su marco CNN 2D, lo que permite la captura de dinámicas temporales tanto a corto como a largo plazo utilizando un mecanismo adaptativo de dos niveles. La Red Piramidal Temporal (TPN), por otro lado, extrae e integra información espacial, temporal y semántica utilizando un reescalado jerárquico, mejorando el rendimiento para tareas con variabilidad temporal. Para TANet y TPN, utilizamos la arquitectura ResNet50 como *backbone*.

Para el conjunto de datos Kinetics 400, utilizamos pesos pre-entrenados disponibles en MMAction2. Para el conjunto de datos EtriActivity3D, realizamos *fine tuning* de las redes utilizando los pesos pre-entrenados de Kinetics 400, entrenando durante 10 épocas con validación cruzada con k=5.

### 3.3. Métodos XAI

Dado que adoptamos redes con arquitecturas variables, optamos por métodos XAI agnósticos del modelo, que generan explicaciones independientemente del modelo subyacente. Específicamente, empleamos una versión adaptada a vídeo de métodos XAI agnósticos del modelo ampliamente utilizados, que se encuentra disponible públicamente[1]. Dichos métodos incluyen adaptaciones de LIME (Ribeiro et al., 2016) (Video LIME), Kernel-SHAP (Lundberg & Lee, 2017) (Video Kernel-SHAP), RISE (Petsiuk

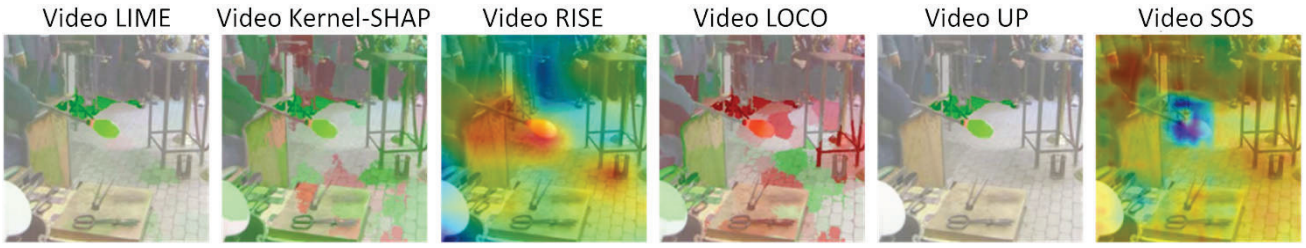



*Figura 1: Ejemplo de explicaciones calculadas usando el modelo TimeSformer y el conjunto de datos Kinetics 400, utilizando los diferentes métodos. Solo se muestra un fotograma del momento de máxima relevancia por método.*

et al., 2018) (Video RISE), *Occlusion sensitivity* (Zeiler & Fergus, 2014) (Video SOS), LOCO (Lei et al., 2018) (Video LOCO) y predictores univariados (Guyon & Elisseeff, 2003) (Video UP). Relacionándolos con las dimensiones de XAI mencionadas anteriormente, estos métodos tienen como objetivo explicar el modelo y se caracterizan por ser *post-hoc*, de alcance local y estáticos en la presentación (Gaya-Morey et al., 2024).

La operación de estos métodos implica cuatro pasos principales: (1) segmentar el vídeo de entrada en regiones que consisten en píxeles de diferentes fotogramas, (2) ocluir estas regiones y pasar el vídeo modificado a través del modelo, donde las predicciones cambian según las regiones ocluidas, (3) resumir la relevancia de cada región para la predicción objetivo y (4) visualizar las explicaciones. Los parámetros exactos utilizados en cada paso dependen del método. La aplicación de estos métodos XAI para explicar la predicción de un modelo para un vídeo dado produce una explicación en forma de vídeo, dentro del cual cada píxel representa la relevancia del píxel correspondiente en el vídeo original. La Figura 1 muestra un ejemplo de explicación utilizando cada método.

### 3.4. Explicaciones de vídeos

Para la evaluación, se seleccionó una muestra aleatoria de 30 vídeos de cada conjunto de datos: Kinetics 400 y EtriActivity3D. Por consistencia, solo se incluyeron vídeos que fueron clasificados correctamente por las tres redes utilizadas en el estudio. Si un vídeo fue mal clasificado por alguna red, fue reemplazado por otro vídeo seleccionado al azar. Para asegurar una comparación justa, se impusieron condiciones iguales para los métodos en la medida de lo posible, como el número de características, muestras y tipo de oclusión.

Cada uno de los 30 vídeos de ambos conjuntos de datos se procesó a través de las tres redes: TimeSformer, TPN y TANet. Para cada predicción, se generaron explicaciones utilizando los seis métodos XAI descritos anteriormente. Esto resultó en 6 participantes × 6 métodos XAI × 3 redes × 2 conjuntos de datos × 30 vídeos por condición =1.080 explicaciones a lo largo del experimento.

Para mejorar la claridad y la interpretabilidad de las explicaciones, solo se retuvo el 30% superior de las regiones más relevantes, filtrando las áreas menos significativas (puede apreciarse más adelante en las Figuras Figura 2 y Figura 5). Además, se aplica estiramiento de histograma para asegurar que las explicaciones utilizaran todo el rango del espectro de color, haciendo que las visualizaciones fueran más claras. Además, se eliminaron los valores de relevancia negativos por dos razones principales: simplificar la información presentada a los usuarios que evalúan las explicaciones y estandarizar las salidas en todos los métodos XAI, ya que no todos los métodos proporcionan puntuaciones de relevancia tanto positivas como negativas.

## 4. Método

Se presentaron explicaciones a los usuarios para evaluar sus preferencias.

### 4.1. Participantes

Seis voluntarios (tres mujeres) del entorno universitario participaron en el estudio. Las edades oscilaron entre 24 y 47 años (media = 34,7; DT = 10,4). Dichos participantes tienen una amplia experiencia tanto en IA como en XAI, con su conocimiento basado en años de investigación especializada y aplicaciones prácticas. Dos de los expertos, los más jóvenes, trabajaron en IA durante al menos tres o cuatro años y han pasado los últimos dos años trabajando en XAI. Los expertos más experimentados tienen una amplia trayectoria tanto en IA como en XAI, habiendo trabajado en esta última área durante un mínimo de cinco años. Además, tres de los expertos centran su investigación en Interacción Persona-Ordenador (IPO). Su investigación abarca un amplio espectro, incluida la visión por computador y el aprendizaje profundo aplicados a problemas de (IPO). Aunque todos los participantes tenían experiencia con IA y XAI, su familiaridad no se extendía a todos los métodos XAI.

### 4.2. Aparato

Se desarrolló una interfaz de usuario para mostrar el vídeo, su clase asociada, la explicación y un mapa de color correspondiente para ayudar a los usuarios en su evaluación (ver Figura 2). La interfaz muestra las 1.080 explicaciones, cada visualización muestra solo una explicación.

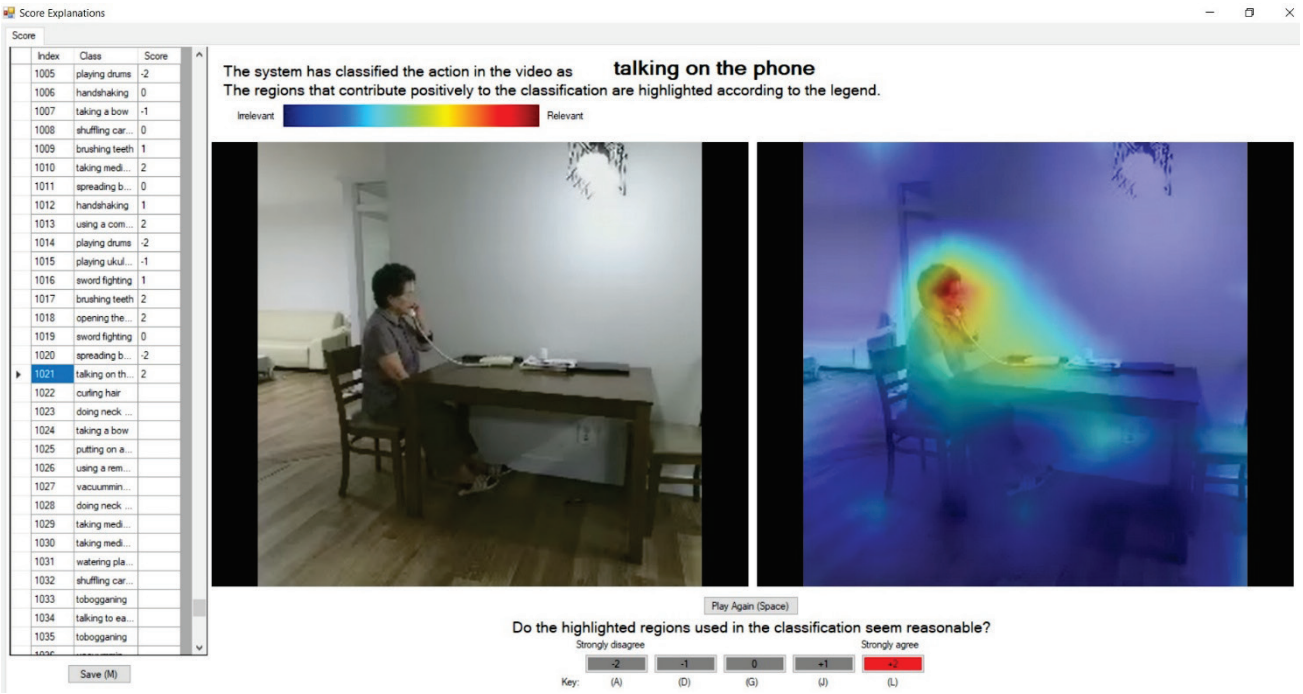



*Figura 2: Interfaz de usuario para puntuar las explicaciones generadas por el método XAI. A la izquierda, una lista desplazable de vídeos con su clase correspondiente y las puntuaciones asignadas por el usuario. En la parte superior, se muestra información pertinente sobre la explicación actual, incluida la clase y el mapa de color que representa la explicación. En el centro, se presenta el vídeo que se está explicando (izquierda) junto con su explicación correspondiente (derecha). En la parte inferior, se presenta al usuario una pregunta y opciones de respuesta, con la respuesta seleccionada en rojo.*

La pregunta planteada a los usuarios durante la evaluación fue: "¿Las regiones resaltadas utilizadas en la clasificación parecen razonables?" Las opciones de respuesta se presentan en una escala de Likert con valores de entre -2 (totalmente en desacuerdo) y +2 (totalmente de acuerdo). Por lo tanto, las puntuaciones positivas resaltan el alineamiento entre las regiones importantes según el método y el usuario, y las negativas, el no alineamiento. La pregunta busca determinar si las regiones resaltadas se alinean con las percepciones de los usuarios al identificar acciones específicas en el vídeo. Para mitigar cualquier posible sesgo, las explicaciones se presentan en orden aleatorio y sin información sobre la red, el conjunto de datos o el método XAI. Esto asegura una evaluación "ciega".

### 4.3. Procedimiento

El estudio se llevó a cabo utilizando un portátil con el programa instalado localmente, que se permitió a los participantes llevar a casa. A cada participante se le encomendó la tarea de evaluar 1.080 explicaciones, un proceso que requirió a cada usuario aproximadamente de 3 a 4 horas. Para acomodar esto, se les dio a los participantes la flexibilidad de pausar y reanudar la evaluación a su conveniencia.

Las explicaciones se presentaron a todos los participantes en el mismo orden, con la siguiente explicación mostrada automáticamente después de evaluar la anterior. Sin embargo, los participantes tuvieron la flexibilidad de poder navegar libremente entre explicaciones, lo que les permitió volver a visitar, reevaluar y actualizar sus puntuaciones según fuera necesario.

Para cada método, se calculó la puntuación media para todos los participantes para cada método XAI. Además, creamos gráficos de barras agregados de las puntuaciones de los participantes por método, conjunto de datos y red y analizamos la significancia estadística de los diferentes factores.

### 4.4. Diseño

El estudio siguió un diseño intrasujetos de 6×3×2 con las siguientes variables independientes y niveles:

- Método XAI (Video RISE, Video Kernel-SHAP, Video LOCO, Video LIME, Video SOS, Video UP)
- Red (TimeSformer, TANet, TPN)
- Conjunto de datos (EtriActivity3D, Kinetics400)

La variable dependiente fue la puntuación de razonabilidad en una escala Likert de 5 puntos de -2 (*totalmente en desacuerdo*) a 2 (*totalmente de acuerdo*).

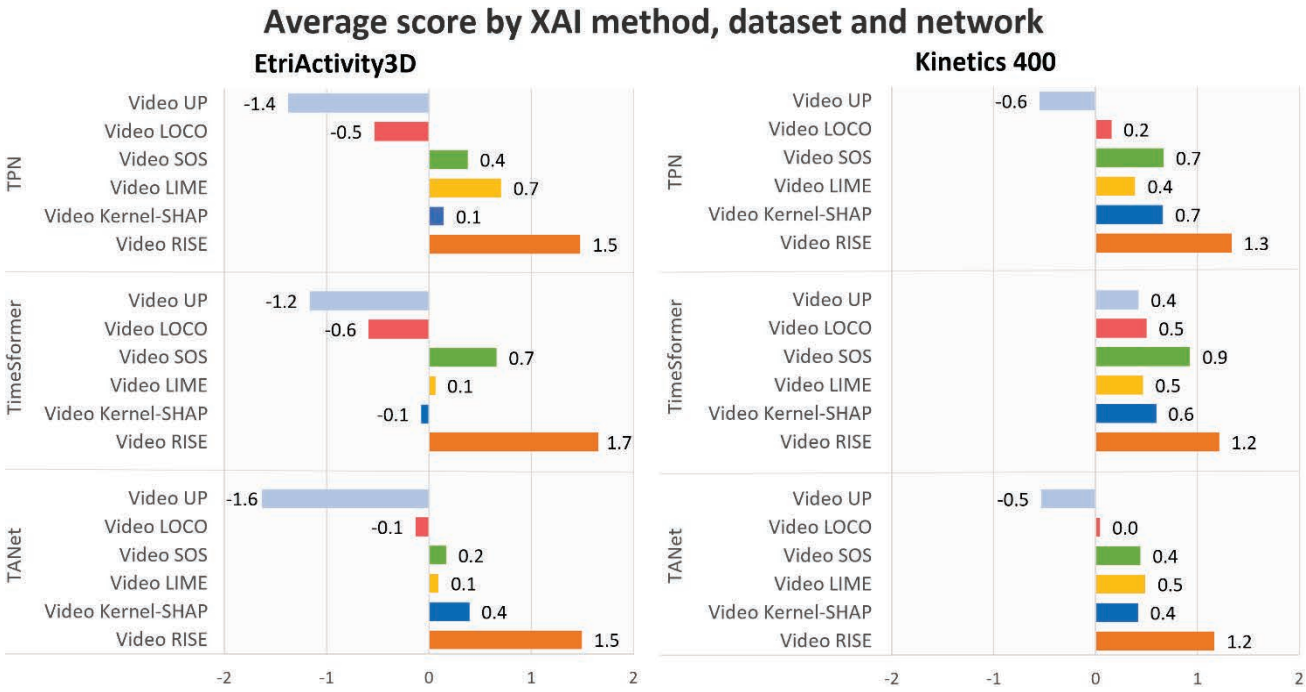



*Figura 3: Promedio de puntuaciones de los usuarios agrupadas por método XAI, conjunto de datos y red.*

El número total de ensayos fue de 6.480 (= 6 participantes × 6 métodos XAI × 3 redes × 2 conjuntos de datos × 30 vídeos por condición).

## 5. Resultados

En esta sección se presentan los resultados de las evaluaciones realizadas por parte de los participantes. La media general sobre las 6.480 explicaciones fue de 0,292. Respecto a la pregunta de interés, las respuestas muestran valores entre 0 (neutral) y 1 (ligeramente de acuerdo), por lo tanto, hubo una tendencia general de los participantes a sentir que las explicaciones de los métodos XAI se inclinaban hacia "razonables". Por método XAI, las medias, de menor a mayor, fueron de -0,806 (Video UP), -0,093 (Video LOCO), 0,368 (Video SOS), 0,356 (Video Kernel-SHAP), 0,540 (Video LIME) y 1,390 (Video RISE). Por red, las medias fueron 0,200 (TANet), 0,389 (TimeSformer) y 0,288 (TPN). Por conjunto de datos, las medias fueron 0,096 (EtriActivity3D) y 0,489 (Kinetics 400). A continuación, se describen múltiples análisis realizados por diferentes combinaciones de estas condiciones. Los resultados de la evaluación de los usuarios se presentan en la Figura 3 y la Figura 4. La Figura 3 ilustra las puntuaciones promedio por método, mientras que la Figura 4 agrega las puntuaciones por conjunto de datos, red y método.

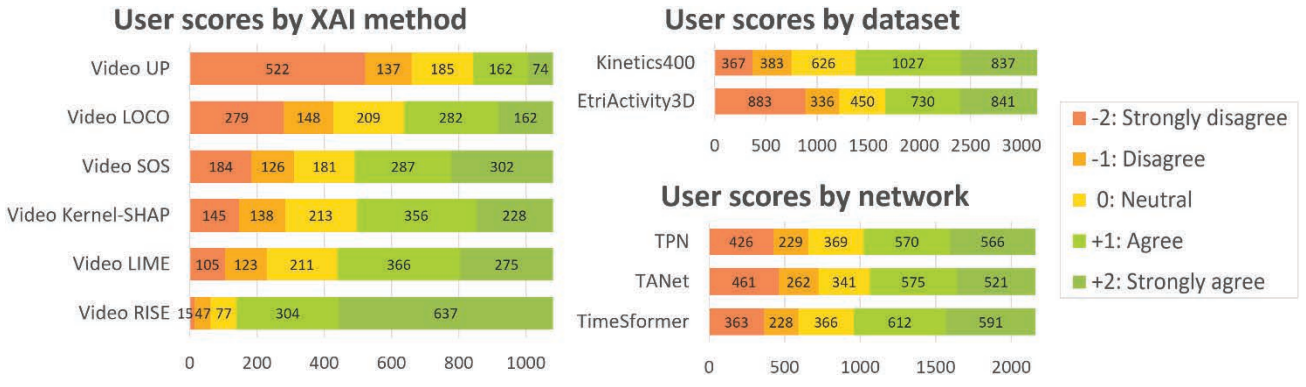



*Figura 4: Recuento de puntuaciones de usuario agrupadas por método XAI, por conjunto de datos y por red.*

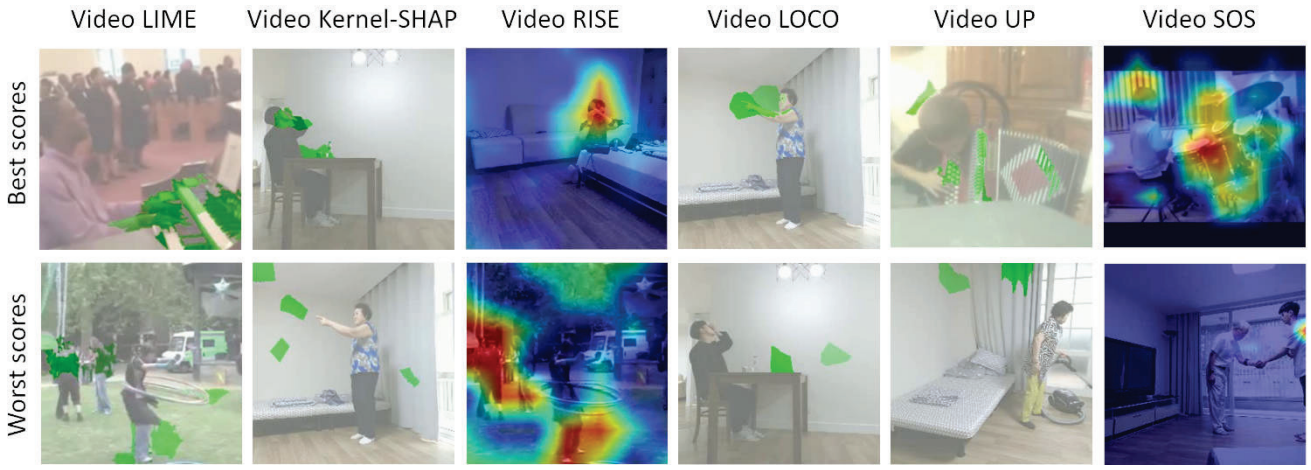



*Figura 5: Imágenes de ejemplo que recibieron las puntuaciones más altas y bajas de los seis expertos. Las columnas representan diferentes métodos XAI, mientras que las filas muestran las imágenes con la puntuación más alta (arriba) y las imágenes con la puntuación más baja (abajo).*

Primero, se realizó un ANOVA de tres vías para evaluar los efectos del conjunto de datos, la arquitectura de la red neuronal y el método XAI en las valoraciones de los usuarios. Se encontraron efectos principales significativos para el conjunto de datos ($F_{1,6444}$=162,83), la red ($F_{2,6444}$=12,49) y el método XAI ($F_{5,6444}$=369,43). Además, se observaron interacciones significativas entre el conjunto de datos y la red ($F_{2,6444}$=10,98), el conjunto de datos y el método XAI ($F_{5,6444}$ = 43,17), la red y el método XAI ($F_{10,6444}$=9,06), y la interacción de tres vías entre el conjunto de datos, la red y el método XAI ($F_{10,6444}$=5,61). En todos los casos, p<0,001. Estos resultados indican que la percepción del usuario no solo depende de factores individuales, sino también de sus combinaciones, con el método XAI mostrando el efecto más fuerte en las valoraciones.

Para evaluar el poder explicativo de diferentes factores, calculamos $R^2$ para tres modelos: uno que incluye todos los factores (conjunto de datos, red y método XAI), uno que considera solo el método XAI y uno que incluye solo el conjunto de datos y la red. El modelo completo alcanzó $R^2$ = 0,273, lo que indica que los factores juntos explican el 27,3% de la varianza en las valoraciones de los usuarios. El modelo que considera solo el método XAI arrojó $R^2$ = 0,208, confirmando que el método XAI es el factor más influyente. Por el contrario, el modelo que incluye solo el conjunto de datos y la red arrojó solo $R^2$ = 0,024, lo que sugiere que estos factores por sí solos contribuyen mínimamente a explicar las valoraciones de los usuarios. Además, el modelo completo tuvo el AIC (21.216) y el BIC (21.460) más bajos, lo que indica el mejor equilibrio entre la bondad de ajuste y la complejidad del modelo.

A continuación, se realizó una prueba *post hoc* de Tukey HSD para analizar las diferencias por pares entre los métodos XAI. Los resultados mostraron diferencias significativas en la mayoría de las comparaciones (p<0,05), excepto entre Video Kernel-SHAP y Video SOS (p=0,999), donde no se encontró ninguna diferencia significativa. El método Video RISE obtuvo consistentemente valoraciones significativamente más altas en comparación con otros métodos, con las mayores diferencias observadas frente a Video UP (diferencia media=2,20; p<0,001). Por el contrario, Video UP recibió valoraciones significativamente más bajas que todos los demás métodos. Estos hallazgos confirman que la elección del método XAI influye fuertemente en las valoraciones de los usuarios. Video RISE exhibe los resultados más favorables, alcanzando una puntuación promedio de 1,39 dentro del rango [-2, 2]. En segundo lugar, se encuentra Video LIME (0,54), seguido de cerca por Video SOS (0,37) y Video Kernel SHAP (0,36). Por el contrario, Video LOCO puntúa mal (-0,09), y Video UP recibe una puntuación de -0,81, la puntuación más baja.

También se realizó una prueba *post hoc* de Tukey HSD para analizar las diferencias por pares entre las tres arquitecturas de redes neuronales. Los resultados revelaron que TimeSformer recibió valoraciones significativamente más altas que TANet (diferencia media=0,19; p<0,001). Sin embargo, las diferencias entre TPN y TANet (diferencia media=0,09; p=0,119) y entre TimeSformer y TPN (diferencia media = 0,10; p = 0,056) no fueron estadísticamente significativas.

## 6. Discusión

La preferencia por Video RISE por parte de los expertos sugiere que colocar regiones importantes sobre la persona que realiza la acción tiene sentido para los usuarios (ver Figura 5, primera fila, tercera columna, explicación para la clase "cepillarse el pelo"). Además, las explicaciones suaves mostradas por Video RISE, sin bordes duros, fueron mejor

valoradas sobre otros métodos. Esta observación plantea la cuestión de si introducir suavidad en otros métodos, como a través de un filtro gaussiano, influiría positivamente en la calidad de la explicación según los usuarios. Si bien Video RISE logró consistentemente resultados superiores en todos los conjuntos de datos y redes, el rendimiento de otros métodos varió dependiendo de estos factores. Por ejemplo, Video UP puntuó aproximadamente un punto más alto en Kinetics 400 que en EtriActivity3D, y Video SOS funcionó mejor en TimeSformer que en otras redes. Esto sugiere que ciertos métodos XAI pueden ser más adecuados para redes neuronales o características de datos específicas.

El conjunto de datos también influyó en las valoraciones de los usuarios. En promedio, las puntuaciones para Kinetics 400 fueron 0,39 puntos más altas que las de EtriActivity3D, con el ANOVA confirmando esta diferencia como significativa ($F_{1,6444}$=162,8; p<0,001). Atribuimos la diferencia a la complejidad del conjunto de datos: Kinetics 400 presenta vídeos más complejos con movimientos de cámara, cortes y una gama más amplia de clases de acción, lo que dificulta la generación de explicaciones. En contraste, EtriActivity3D ofrece un contexto más simple para identificar regiones importantes para la clasificación, lo que probablemente influyó en las puntuaciones de los usuarios.

Con respecto a la selección de la red, observamos diferencias significativas en las puntuaciones promedio de los usuarios entre dos modelos: TimeSformer (puntuación promedio=0,39) y TANet (puntuación promedio=0,20). Sin embargo, no se encontró ninguna diferencia significativa entre TPN (puntuación promedio=0,29) y cualquiera de los otros dos. Esto demuestra que, incluso cuando se entrena en condiciones idénticas, las diferencias de arquitectura entre modelos impactan las evaluaciones de los usuarios. Por ejemplo, las explicaciones con Video UP y Video SOS recibieron consistentemente puntuaciones más altas cuando se generaron para TimeSformer en comparación con las otras dos redes, como se muestra en la Figura 3. En consecuencia, para garantizar una evaluación justa de los métodos XAI, los experimentos deben incluir múltiples redes que representen diversas arquitecturas.

## 7. Limitaciones del estudio

Una limitación de este estudio es el tamaño de la muestra de participantes. Sin embargo, el acuerdo unánime entre los participantes tanto en las mejores como en las peores explicaciones fortalece nuestra confianza en los hallazgos. La Figura 5 presenta ejemplos de explicaciones que recibieron por unanimidad las puntuaciones más altas y bajas.

Otra posible amenaza a la validez del estudio es la muestra de métodos XAI, redes, datasets y actividades elegidas. Aunque se ha incluido relativa variedad a cada uno de estos aspectos, al haberse constatado que algunos de ellos tienen un impacto directo en las puntuaciones del usuario (por ejemplo el dataset y la red), cabe plantearse la necesidad de incluir mayor variedad en futuros estudios, e incluso explorar qué características concretas están influyendo en las puntuaciones de los usuarios.

Finalmente, hay que mencionar que el enfoque de este estudio es sobre usuarios expertos en IA, capaces de entender explicaciones en forma de mapas de calor y sus implicaciones para con las redes neuronales usadas. Por tanto, se ha dejado para trabajo futuro la comprobación del impacto que pueda tener este factor de conocimiento del ámbito sobre las puntuaciones.

## 8. Conclusión

Aunque existen numerosos métodos XAI para generar explicaciones, seleccionar el método más adecuado sigue siendo un reto tanto para investigadores como para profesionales. Este estudio marca un paso adelante en la comprensión de cómo los usuarios perciben seis métodos XAI conocidos (incluidos LIME, SHAP o RISE) cuando se adaptan al dominio del vídeo. Al aplicar explicaciones en diversos conjuntos de datos y redes, se analiza la influencia de esos factores. Sorprendentemente, y aunque la muestra de expertos es pequeña, hubo consenso: Video RISE fue el preferido por los participantes, mientras que Video UP recibió las puntuaciones más bajas.

Los estudios de usuario para evaluar los métodos XAI son esenciales para obtener información sobre cómo los usuarios interactúan e interpretan las explicaciones de los sistemas de IA. Este conocimiento puede guiar las decisiones técnicas basadas en las preferencias de explicabilidad de los usuarios y ayudar a elegir un método XAI apropiado para aplicaciones del mundo real.

Sin embargo, los estudios consumen mucho tiempo y son costosos, ya que requieren gran cantidad de recursos para recopilar datos significativos. Para acelerar el proceso de evaluación, las métricas automáticas como el área bajo la curva (AUC) pueden ofrecer formas más eficientes de evaluar los métodos XAI. Sin embargo, persiste un debate sobre si el rendimiento de los métodos XAI a través de métricas objetivas debe tener prioridad sobre las preferencias del usuario al determinar su efectividad o aplicación. No obstante, es fundamental probar qué tan bien se alinean las métricas automáticas con la perspectiva del usuario. Cerrar

esta brecha asegurará que el proceso de evaluación siga siendo efectivo y representativo de las experiencias reales del usuario.

El trabajo futuro implicará evaluaciones con una muestra más grande de usuarios para validar y probar aún más nuestros resultados. Además, incorporar una mayor diversidad de participantes, como variaciones en el conocimiento de IA, la edad y otros datos demográficos, proporcionará una comprensión más profunda de los métodos XAI desde una perspectiva humana.

## Agradecimientos

Este trabajo forma parte del proyecto PID2023-149079OB-I00 (EXPLAINME), financiado por MICIU/AEI/10.13039/501100011033/ y FEDER (UE), y del proyecto PID2022-136779OB-C32 (PLEISAR), financiado por MICIU/AEI/10.13039/501100011033/ y FEDER (UE).

F. X. Gaya-Morey contó con el apoyo de una beca FPU del Ministerio de Fondos Europeos, Universidad y Cultura del Gobierno de las Islas Baleares.

## Referencias

Aechtner, J., Cabrera, L., Katwal, D., Onghena, P., Valenzuela, D. P., & Wilbik, A. (2022). Comparing User Perception of Explanations Developed with XAI Methods. *Proceedings of the IEEE International Conference on Fuzzy Systems – FUZZ-IEEE '22*, 1-7. https://doi.org/10.1109/FUZZ-IEEE55066.2022.9882743

Alqaraawi, A., Schuessler, M., Weiss, P., Costanza, E., & Berthouze, N. (2020). Evaluating saliency map explanations for convolutional neural networks: A user study. *Proceedings of the 25th International Conference on Intelligent User Interfaces – IUI '20*, 275-285. https://doi.org/10.1145/3377325.3377519

Arya, V., Bellamy, R. K. E., Chen, P.-Y., Dhurandhar, A., Hind, M., Hoffman, S. C., Houde, S., Liao, Q. V., Luss, R., Mojsilović, A., Mourad, S., Pedemonte, P., Raghavendra, R., Richards, J. T., Sattigeri, P., Shanmugam, K., Singh, M., Varshney, K. R., Wei, D., & Zhang, Y. (2020). AI Explainability 360: An Extensible Toolkit for Understanding Data and Machine Learning Models. *Journal of Machine Learning Research*, *21*(130), 1-6. http://jmlr.org/papers/v21/19-1035.html

Barda, A. J., Horvat, C. M., & Hochheiser, H. (2020). A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare. *BMC Medical Informatics and Decision Making*, *20*(1). https://doi.org/10.1186/s12911-020-01276-x

Bertasius, G., Wang, H., & Torresani, L. (2021). Is Space-Time Attention All You Need for Video Understanding? *Proceedings of the 38th International Conference on Machine Learning Research – PMLR '21*, *139*, 813-824. https://proceedings.mlr.press/v139/bertasius21a/bertasius21a-supp.pdf

Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K. E., & Dugan, C. (2019). Explaining models: An empirical study of how explanations impact fairness judgment. *Proceedings of the 24th International Conference on Intelligent User Interfaces – IUI '19*, 275-285. https://doi.org/10.1145/3301275.3302310

Ehsan, U., Passi, S., Liao, Q. V., Chan, L., Lee, I.-H., Muller, M., & Riedl, M. O. (2024). The Who in XAI: How AI Background Shapes Perceptions of AI Explanations. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems – CHI '24*, 316.1-316.32. https://doi.org/10.1145/3613904.3642474

Ehsan, U., Wintersberger, P., Liao, Q. V., Watkins, E. A., Manger, C., Daumé III, H., Riener, A., & Riedl, M. O. (2022). Human-Centered Explainable AI (HCXAI): Beyond Opening the Black-Box of AI. *Extended Abstracts of the ACM SIGCHI Conference on Human Factors in Computing Systems – CHI EA '22*. https://doi.org/10.1145/3491101.3503727

Gaya-Morey, F. X., Buades-Rubio, J. M., MacKenzie, I. S., & Manresa-Yee, C. (2024). *REVEX: A Unified Framework for Removal-Based Explainable Artificial Intelligence in Video*. https://doi.org/10.48550/arXiv.2401.11796

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, *3*(null), 1157-1182. https://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf

Heimerl, A., Weitz, K., Baur, T., & Andre, E. (2020). Unraveling ML Models of Emotion with NOVA: Multi-Level Explainable AI for Non-Experts. *IEEE Transactions on Affective Computing*, *1*(1), 1-13. https://doi.org/10.1109/TAFFC.2020.3043603

Jang, J., Kim, D., Park, C., Jang, M., Lee, J., & Kim, J. (2020). ETRI-Activity3D: A Large-Scale RGB-D Dataset for Robots to Recognize Daily Activities of the Elderly. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems – IROS '20*, 10990-10997. https://doi.org/10.1109/IROS45743.2020.9341160

Kaplan, S., Uusitalo, H., & Lensu, L. (2024). A unified and practical user-centric framework for explainable artificial intelligence. *Knowledge-Based Systems*, *283*, 111107. https://doi.org/10.1016/j.knosys.2023.111107

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., & Zisserman, A. (2017). *The Kinetics Human Action Video Dataset*. https://doi.org/10.48550/arXiv.1705.06950

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-Free Predictive Inference for Regression. *Journal of the American Statistical Association*, *113*(523), 1094-1111. https://doi.org/10.1080/01621459.2017.1307116

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing Design Practices for Explainable AI User Experiences. *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems – CHI '20*, 1-15. https://doi.org/10.1145/3313831.3376590

Liao, Q. V., & Varshney, K. R. (2022). *Human-centered explainable AI (XAI): From algorithms to user experiences*. https://doi.org/10.48550/arXiv.2110.10790

Liu, Z., Wang, L., Wu, W., Qian, C., & Lu, T. (2021). TAM: Temporal Adaptive Module for video recognition. *Proceedings of the IEEE/CVF International Conference on Computer Vision – ICCV '21*, 13688-13698. https://doi.org/10.1109/ICCV48922.2021.01345

Lopes, P., Silva, E., Braga, C., Oliveira, T., & Rosado, L. (2022). XAI Systems Evaluation: A Review of Human and Computer-Centred Methods. *Applied Sciences*, *12*(19). https://doi.org/10.3390/app12199423

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems – NIPE '17*, 4768-4777. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

Manresa-Yee, C., Ramis, S., Gaya-Morey, F. X., & Buades, J. M. (2024). Impact of Explanations for Trustworthy and Transparent Artificial Intelligence. *Proceedings of the XXIII International Conference on Human Computer Interaction– Interacción '23*. https://doi.org/10.1145/3612783.3612798

Miller, T. (2019). Explanation in Artificial Intelligence: Insights From the Social Sciences. *Artificial Intelligence*, *267*(C), 1-38. https://doi.org/10.1016/j.artint.2018.07.007

Mohseni, S., Zarei, N., & Ragan, E. D. (2018). A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Transactions on Interactive Intelligent Systems*, *11*, 24:1–24:45. https://doi.org/10.1145/3387166

OpenMMLab. (2020). *OpenMMLab's Next Generation Video Understanding Toolbox and Benchmark*.

Petsiuk, V., Das, A., & Saenko, K. (2018). RISE: Randomized Input Sampling for Explanation of Black-box Models. *Proceedings of the British Machine Vision Conference – BMVC '18, Newcastle, UK*, 1-151. https://doi.org/10.48550/arXiv.1806.07421

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). «Why Should I Trust You?»: Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD '16*, 1135-1144. https://doi.org/10.1145/2939672.2939778

Ridley, M. (2025). Human-centered explainable artificial intelligence: An Annual Review of Information Science and Technology (ARIST) paper. *Journal of the Association for Information Science and Technology*, *76*(1), 98-120. https://doi.org/https://doi.org/10.1002/asi.24889

Rong, Y., Leemann, T., Nguyen, T.-T., Fiedler, L., Qian, P., Unhelkar, V., Seidel, T., Kasneci, G., & Kasneci, E. (2024). Towards human-centered explainable AI: A survey of user studies for model explanations . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, *46*(04), 2104-2122. https://doi.org/10.1109/TPAMI.2023.3331846

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Proceedings of the International Conference on Computer Vision – ICCV '17*, 618-626. https://doi.org/10.1109/ICCV.2017.74

Shneiderman, B. (2020). Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495-504. https://doi.org/10.1080/10447318.2020.1741118

Szymanski, M., Millecamp, M., & Verbert, K. (2021). Visual, textual or hybrid: The effect of user expertise on different explanations. *Proceedings of the 26th International Conference on Intelligent User Interfaces – IUI '21*, 109-119. https://doi.org/10.1145/3397481.3450662

Wells, L., & Bednarz, T. (2021). Explainable AI and Reinforcement Learning: A systematic review of current approaches and trends. *Frontiers in Artificial Intelligance*, *4*, 1-15. https://doi.org/10.3389/frai.2021.550030

Yang, C., Xu, Y., Shi, J., Dai, B., & Zhou, B. (2020). Temporal Pyramid Network for Action Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition – CVPR '20*, 591-600. https://doi.org/10.1109/CVPR42600.2020.00067

Zeiler, M. D., & Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. *Proceedings of the 13th European Conference on Computer Vision – ECCV '14 (LNCS 8689)*, 818-833. https://doi.org/10.1007/978-3-319-10590-1_53