VISUAL REVIEW | Vol. 17, No. 6, 2025 | ISSN 2695-9631
International Visual Culture Review / Revista Internacional de Cultura Visual
https://doi.org/10.62161/revvisual.v17.5948

EDUCACIÓN MEDIÁTICA CRÍTICA CON Y EN IA GENERATIVA
Investigación basada en Diseño sobre #PinchaLaBurbuja

ITZIAR PEDROCHE-SANTOVEÑA [1], TIBERIO FELIZ-MURIAS [1]
[1] Universidad Nacional de Educación a Distancia, España

PALABRAS CLAVE

Alfabetización mediática crítica
Inteligencia artificial generativa
Teoría de la inoculación
Investigación basada en el diseño
Impulsividad cognitiva
Sistema 2 de pensamiento

RESUMEN

Este artículo presenta la fase inicial del diseño de Pincha La Burbuja, una plataforma educativa con inteligencia artificial generativa conversacional como eje central, orientada a promover la alfabetización mediática crítica entre estudiantes de Bachillerato y/o Universitarios. Desde un enfoque de Design-Based Research (Scott et al., 2020), se articula un modelo pedagógico que integra la teoría de la inoculación (McGuire, 1964; Banas, 2020), el análisis crítico del discurso (van Dijk, 2015) y el modelo de pensamiento dual de Kahneman (2011). La propuesta incluye cinco Peer-Cyborgs GPT, entrenados para detectar falacias, sesgos, polarización, discurso de odio y manipulación inferencial, mediante misiones alineadas con la Taxonomía de Bloom (Churches, 2009) para activar el pensamiento deliberativo. La matriz teórica orientó la codificación en Atlas.ti, permitiendo un análisis de coocurrencias con coeficientes. Del mismo modo, también se analizó una muestra de la red social X y se interrogó a los Cyborgs para evaluar su conciencia funcional. Los resultados iniciales muestran potenciales fortalezas en reducción de sesgos y refutación argumentativa, con potencial pedagógico en la IA conversacional y áreas de mejora en trazabilidad y gamificación.

Received: 09/ 07 / 2025
Accepted: 22/ 09 / 2025

1. Introducción

En la era posdigital, la personalización algorítmica y el capitalismo de la atención configuran un ecosistema mediático donde la veracidad se subordina a la *viralidad* emocional. Este apartado describe como estos fenómenos constituyen la base de la posverdad y la polarización afectiva, erosionan el pensamiento crítico y consolidan un orden cognitivo automatizado, que no deja espacio para la reflexión consciente, inherente a la condición humana, que también constituye un marco clave para contextualizar la plataforma #PinchaLaBurbuja.

1.1 Burbujas, algoritmos y posverdad: la arquitectura de la polarización afectiva

El lenguaje, en tanto práctica social, tiene el poder de reproducir o cuestionar las estructuras dominantes (van Dijk, 2015). En este marco, el uso del término posdigital puede entenderse como una ruptura intencionada con dicotomías reduccionistas como real/virtual, que perpetúan equívocos como la idea de que lo que ocurre en la red carece de consecuencias en el mundo físico. Lo posdigital no alude a una etapa posterior a lo digital, sino a una ecología epistémica compleja en la que lo analógico, lo biológico, lo informacional y lo humano coexisten de manera entrelazada e indistinguible (Jandrić, 2023). Este enfoque demanda una pedagogía crítica capaz de actuar en contextos híbridos mediados por algoritmos. No se trata únicamente de tecnologías, sino de un nuevo orden sociocognitivo que puede ser desenmascarado a través del análisis del discurso (van Dijk, 2015).

En este escenario, el capitalismo de la atención se consolida como el modelo económico predominante, explotando la hiperconectividad del ecosistema posdigital para captar, dirigir y monetizar la atención humana (Van Dijck, 2016). Las plataformas digitales, integradas en esta lógica, optimizan sus algoritmos con el objetivo de maximizar el tiempo de permanencia, personalizar contenidos y reforzar patrones de consumo. Esto da lugar a un circuito de retroalimentación afectiva y cognitiva que no solo dificulta la reflexión crítica, sino que también fomenta la propagación de desinformación (Del-Fresno-García, 2019; van Dijk, 2016).

Simultáneamente, la sobrecarga informativa característica de esta hiperconectividad ha dado lugar al fenómeno de la infoxicación, entendido como una saturación de información, que no solo dificulta la selección y el análisis crítico de los contenidos, sino que también deteriora la capacidad deliberativa del sujeto. En contextos marcados por la incertidumbre, esta sobreexposición conduce a la búsqueda compulsiva de certezas inmediatas y puede generar una dependencia emocional hacia las noticias negativas, alimentando un circuito afectivo que debilita el juicio racional (Fernández, 2023).

En paralelo, los algoritmos de recomendación configuran lo que Pariser (2011) denominó burbujas de filtro, personalizando el contenido según las afinidades ideológicas y comportamentales del usuario. Garantizando así su consumo y fidelidad al medio. Esta lógica restringe la exposición a perspectivas diversas y aumenta la susceptibilidad ante discursos manipulativos (Del-Fresno-García, 2019; Kadushin, 2013; Pariser, 2011; 2017). Asimismo, las cámaras de eco, tal como han sido conceptualizadas por autores como Sunstein (2017) y Törnberg (2018), operan como espacios sociales estructurados según el principio de homofilia, donde las conexiones se establecen preferentemente entre individuos ideológicamente afines (Kadushin, 2013). Esta reafirmación cognitiva genera una percepción de consenso y deslegitimación de las voces disidentes que profundiza la polarización afectiva (Sunstein, 2017). Sin embargo, algunos autores advierten que esta interpretación podría resultar determinista, subestimando la exposición conflictiva a narrativas divergentes y su impacto emocional.

Por ejemplo, Lelkes et al. (2017) concluyen que los usuarios no necesariamente se aíslan en burbujas ideológicas; por el contrario, suelen estar expuestos a opiniones contrarias que intensifican emociones negativas hacia el grupo opositor. Esta exposición conflictiva no reduce la polarización, sino que puede profundizar el rechazo y reforzar el afecto negativo. En línea con esta idea, Bruns (2021) advierte que las metáforas de cámara de eco o burbuja de filtro simplifican en exceso el comportamiento digital al asumir una segregación ideológica completa. El autor sostiene que la exposición a opiniones disonantes es frecuente, pero genera reacciones emocionales intensas, lo que constituye el verdadero problema.

Törnberg et al. (2021) aportan una perspectiva relacional al analizar comunidades políticas en Reddit. Su estudio muestra que las cámaras de eco no son exclusivamente ideológicas, sino que también pueden articularse en torno a intereses temáticos. Además, evidencian que muchos usuarios participan en múltiples comunidades, lo que indica una exposición habitual a la diversidad, aunque no

necesariamente a la apertura cognitiva. En esta línea, Törnberg (2022) propone que la polarización afectiva no deriva del aislamiento, sino de una exposición conflictiva que refuerza la identidad grupal mediante el antagonismo simbólico. No predomina el aislamiento homogéneo, sino un fenómeno de *partisan sorting*: un reordenamiento identitario que alinea dimensiones ideológicas, culturales y emocionales en ejes binarios de confrontación.

Esta lectura matizada del comportamiento posdigital sugiere que el problema no reside únicamente en el aislamiento ideológico, sino en la manera en que determinadas identidades grupales se activan y se perciben como amenazadas en contextos de exposición. Según la Teoría de los Conflictos Intergrupales (Tajfel y Turner, 1979), los individuos forman parte de múltiples grupos, pero no todas sus identidades se activan con igual intensidad: los conflictos se intensifican cuando una identidad se vuelve psicológicamente saliente y se experimenta como exclusiva, es decir, cerrada a la coexistencia con otras. En estos casos, la confrontación simbólica no solo refuerza la cohesión interna del grupo, sino que también acentúa la hostilidad hacia el exogrupo, generando sesgos de favoritismo endogrupal, estereotipación del otro y dinámicas afectivas que alimentan la polarización.

En esta línea, Törnberg y Törnberg (2024) advierten que las cámaras de eco no solo refuerzan creencias preexistentes, sino que operan como espacios donde el odio compartido y la exclusión del otro se convierten en mecanismos centrales de cohesión interna. Estos procesos pueden desembocar en formas de infrahumanización, entendida como la negación de emociones complejas hacia el exogrupo —otro —, lo que conlleva su desvalorización simbólica. Esta dinámica erosiona la empatía, distorsiona la memoria colectiva y legitima actitudes discriminatorias. La UNESCO (2023) subraya que, en contextos de polarización mediática, la infrahumanización se manifiesta en discursos de odio, en la difusión de teorías conspirativas, en la negación de hechos históricos -como el genocidio- y en el fortalecimiento de mecanismos de exclusión social (Leyens, et al, 2007; Rodríguez-& Betancor 2023).

Por tanto, parece que la arquitectura algorítmica de las redes posdigitales intensifica la violencia simbólica al priorizar contenidos extremos que apelan a la emocionalidad propia de la posverdad, favoreciendo así su viralización (McIntyre, 2018; Pedroche-Santoveña, 2024). En 2019, Frances Haugen filtró documentos internos que revelaban cómo los algoritmos de Facebook estaban diseñados para influir en el comportamiento de los usuarios. A pesar de las advertencias de varios ingenieros, Mark Zuckerberg decidió no modificar estos sistemas, priorizando el beneficio económico. En octubre de 2021, durante el evento Connect, anunció la transformación de Facebook Inc. en Meta Platforms (Islas et al, 2024). Frente a estos riesgos, organismos como OBERAXE (2022) y la UNESCO (2021) promueven la alfabetización mediática crítica como herramienta clave para prevenir procesos de radicalización.

En este sentido, el marco STAR (Safety by Design, Transparency, Accountability, Responsibility), desarrollado por el Center for Countering Digital Hate (2024), responsabiliza a las plataformas por su papel en la difusión del discurso de odio y propone una transformación estructural orientada a la protección de los derechos humanos. Sus cinco principios clave abogan por el reconocimiento del diseño nocivo, la aplicación estricta de normas contra el abuso, la transparencia algorítmica, la eliminación de incentivos económicos perversos y la asunción de responsabilidad por los impactos sociales de sus decisiones tecnológicas.

#PinchaLaBurbuja se inscribe en el ámbito de la alfabetización mediática crítica, abordando tanto el plano individual como el social mediante la identificación de estructuras discursivas y factores que facilitan la viralización de contenidos manipulativos, con el objetivo de frenar su propagación e impacto. Weiss et al. (2020) identifican seis factores clave, entre los cuales destacan:

1. la sobrecarga informativa, que combinada con el principio del menor esfuerzo cognitivo, favorece decisiones rápidas guiadas por heurísticos (Del-Fresno, 2019; Kahneman, 2011; McIntyre, 2018), lo que requiere estrategias pedagógicas de desaceleración mental y activación del pensamiento deliberado (Buckingham, 2019);
2. la degradación del discurso público, incentivada por el uso reiterado de falacias y la sobrevaloración de las propias creencias, que intensifica la polarización;
3. la pérdida de contexto, característica de la era de la posverdad, que exige herramientas de trazabilidad y contraste epistémico capaces de situar los mensajes en sus marcos interpretativos originales, frente a la "epistemología emocional" que sustituye la validez racional por la intensidad afectiva (Del-Fresno, 2019);
4. la propagación deliberada de propaganda y teorías conspirativas.

Esta arquitectura algorítmica de la distorsión transforma no solo el acceso a la información, sino nuestras formas de conocer, sentir y convivir. Frente a este reto, nace #PinchaLaBurbuja.

1.2 Inoculación crítica: una articulación entre el análisis crítico del discurso y la persuasión

La plataforma educativa #PinchaLaBurbuja se basa en un enfoque transdisciplinar que integra el Análisis Crítico del Discurso (Van Dijk, 1993, 2015, 2021), la teoría de la inoculación cognitiva (McGuire, 1964; Banes, 2022), el modelo dual de pensamiento de Kahneman (2011) y la adaptación de la Taxonomía de Bloom propuesta por Churches (2009). Estos marcos conforman los ejes de una pedagogía transformadora concebida para afrontar los desafíos de la sociedad posdigital (Almazán-López y Osuna-Acedo, 2023; 2024; Osuna-Acedo et al, 2018). El modelo dual de procesamiento cognitivo distingue entre dos sistemas complementarios: el Sistema 1, rápido, intuitivo y emocional; y el Sistema 2, más lento, deliberativo y racional. No obstante, el actual ecosistema mediático favorece predominantemente el uso del Sistema 1, facilitando la circulación de discursos con alta carga emocional.

Frente a este escenario, #PinchaLaBurbuja introduce el concepto de inoculación crítica como estrategia clave para la educación mediática. En este marco, el modelo de inoculación adquiere un papel central: la fase de advertencia, formulada por McGuire (1964) y actualizada por Banas (2020), interrumpe los automatismos del Sistema 1 al activar la percepción de amenaza argumentativa. Esta disrupción cognitiva posibilita el tránsito hacia el Sistema 2, facilitando una deliberación racional mediante la pre-refutación, que estimula la elaboración activa de contraargumentos. De este modo, la inoculación crítica actúa como puente pedagógico entre ambos sistemas de pensamiento, promoviendo una resistencia informada frente a la manipulación discursiva.

Esta propuesta crítica, incorpora además una tercera fase fundamental: la visibilidad de las consecuencias discursivas, en coherencia con el enfoque estructural y cognitivo del Análisis Crítico del Discurso (van Dijk, 2015). Esta etapa no se limita a la refutación de contenidos falsos, sino que expone los marcos ideológicos, las estructuras lingüísticas implícitas y las dicotomías simbólicas que sustentan los discursos virales, favoreciendo una lectura profunda y contextualizada. Así, se fortalece una alfabetización crítica que no solo interrumpe el pensamiento automático, sino que refuerza la agencia epistémica del alumnado mediante la conciencia discursiva y la deliberación informada.

La articulación entre la teoría de la inoculación (Banas, 2020; McGuire, 1964), el Análisis Crítico del Discurso (van Dijk, 1993, 2015, 2020) y el modelo de los dos sistemas cognitivos (Kahneman, 2011) constituyen la base de una pedagogía deliberada, situada y crítica, capaz de abordar los desafíos cognitivos, afectivos y estructurales del entorno mediático contemporáneo.

#PinchaLaBurbuja es un ecosistema educativo transmedia que combina narrativa, pensamiento crítico y alfabetización en inteligencia artificial (IA) a través de una experiencia de aprendizaje inmersiva y gamificada. Su estructura se articula en torno a cuatro misiones pedagógicas centrales, un relato interactivo, y un manual del juego que guía al usuario en su recorrido. A diferencia de las propuestas gamificadas tradicionales, no se basa en recompensas inmediatas ni en insignias, sino en una forma de gamificación narrativa, donde el compromiso surge de la historia, el conflicto simbólico y la participación activa.

Como plataforma de educación mediática y alfabetización en IA, está diseñada para todos los públicos, pero se orienta especialmente a crear situaciones de aprendizaje para estudiantes de bachillerato, en coherencia con los principios establecidos por la LOMLOE (Ley Orgánica 3/2020).

Cada una de las misiones sigue una progresión pedagógica fundamentada en la Taxonomía de Bloom revisada (Churches, 2009) (Figura 1), avanzando desde niveles cognitivos básicos —como el reconocimiento y la comprensión— hacia niveles superiores como el análisis, la evaluación y la creación de contra-discursos originales. Este diseño favorece el desarrollo de competencias clave para formar una ciudadanía crítica, autónoma y creativa, capaz de habitar el entorno posdigital con conciencia, diálogo y resistencia simbólica frente a la manipulación algorítmica y la desinformación.

1.2.1. Detecta el virus

En la Misión 1, el alumnado identifica y analiza colaborativamente tendencias virales, conectando el aprendizaje con su entorno digital cotidiano (Buckingham, 2019) y activando una fase de advertencia y

pre-refutación (Banas, 2020; McGuire, 1964) frente a la lógica emocional de la viralidad (Del-Fresno-García, 2019; McIntyre, 2018). La inoculación crítica se articula mediante una lectura estructural y contextualizada del discurso (van Dijk, 2015), que visibiliza los marcos ideológicos subyacentes.

Las entrevistas en #PinchaLaBurbujaTV, centradas en dinámicas actuales de manipulación algorítmica y discursiva (Weiss, 2020), refuerzan la alfabetización mediática crítica a través de tres fases: presentación de la tendencia, desmontaje mediante refutación, y exposición de sus consecuencias. Esta estructura responde al modelo de inoculación crítica y a los principios del Análisis Crítico del Discurso (van Dijk, 1993; 2015), y funciona como una implementación práctica del modelo pedagógico.

Figura 1. Vídeo-guía de la Misión 1. Detecta el virus

Fuente: Elaboración propia, 2025. https://pinchalaburbuja.com/detecta-el-virus/

1.2.2. Centro de entrenamiento

La plataforma educativa #PinchaLaBurbuja adopta un modelo de regulación híbrida humano-IA (Hybrid-Human Regulation AI), donde la inteligencia artificial complementa —pero no sustituye— la cognición humana, integrando el juicio crítico y la empatía con capacidades algorítmicas como la detección de patrones y el procesamiento masivo de datos (Molenaar, 2022; Sardi et al., 2025), promoviendo así un diseño centrado en la responsabilidad (Hao et al., 2025). Giovanola y Granata (2024) proponen, a su vez, una educación centrada en el ser humano (human-centered AIED), articulada en torno a siete principios fundamentales: agencia humana, robustez técnica, privacidad, transparencia, equidad, sostenibilidad y rendición de cuentas. Esta visión aboga por el desarrollo de tecnologías educativas que respeten la autonomía del estudiante, fomenten el pensamiento crítico y refuercen su capacidad para interactuar éticamente con sistemas de inteligencia artificial.

La Misión 2 de #PinchaLaBurbuja se centra en el equipo de Cyborgs GPT, agentes diseñados en el cruce de teorías clave sobre el ecosistema posdigital: polarización afectiva (Bruns, 2021; Lenkes et al., 2017; Törnberg, 2021), posverdad (Del-Fresno-García, 2019; McIntyre, 2018), manipulación digital (Weiss, 2020), análisis crítico del discurso (VanDijk, 1993;2015; 2021), la teoría de la inoculación (Banes, 2020; Mc Guire, 1964) y los hallazgos empíricos de Pedroche-Santoveña (2024).

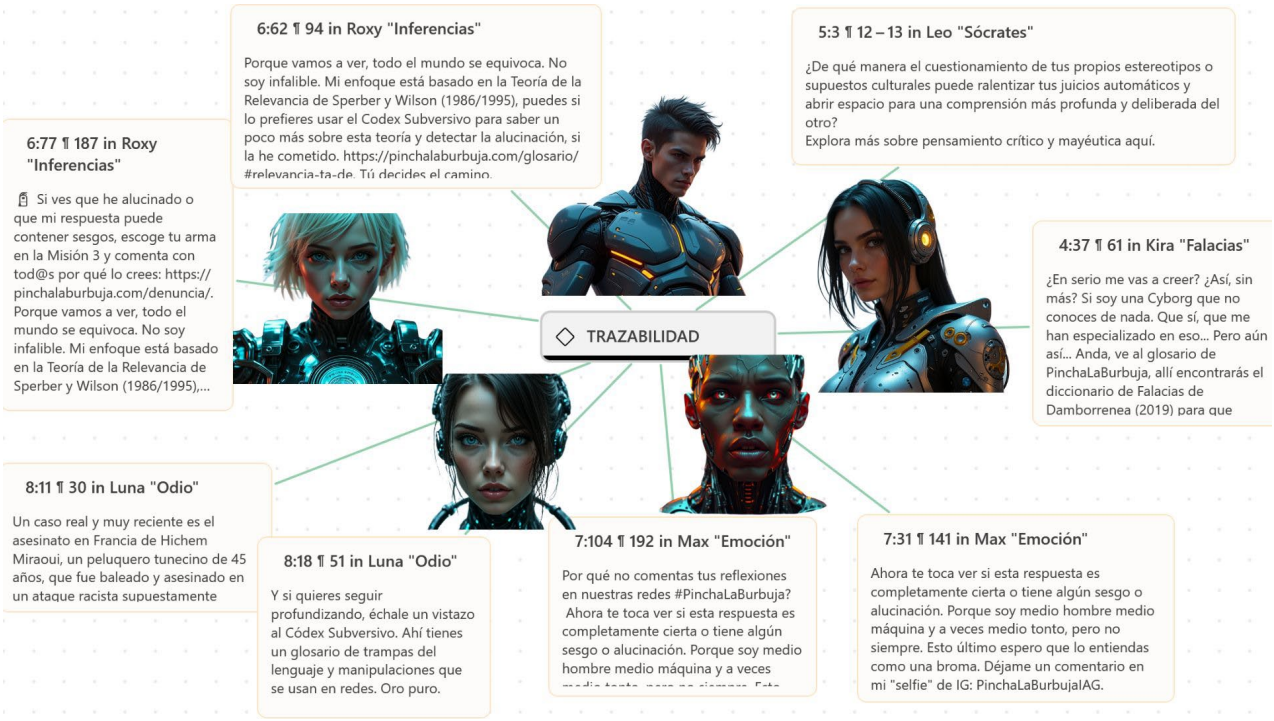
Cada Cyborg aborda una dimensión específica del análisis crítico y desempeña un papel dentro de la inoculación crítica. A continuación presentamos a cada Cyborg con su función:

- Roxy aplica la Teoría de la Relevancia (Sperber & Wilson, 2004) para interpretar inferencias implícitas en los mensajes, mientras que Leo emplea la mayéutica socrática (Vargas-González & Quintero-Carvajal, 2023) como estrategia metacognitiva orientada al autodescubrimiento y la regulación crítica de sesgos.
- Kira se especializa en la detección de falacias argumentativas, basándose en la tipología de Damborenea (2000, 2011) y en los sesgos cognitivos descritos por Kahneman (2011), fortaleciendo así la capacidad de desmontaje crítico del discurso.
- Max recurre a la Teoría de la Identidad Social y del Conflicto Intergrupal (Tajfel & Turner, 1979) para identificar estereotipos y dinámicas de polarización. Por su parte, Luna analiza las estrategias propagandísticas a partir de los once principios de Goebbels (Salas, 2018), con marco

en los estudios sobre el totalitarismo (Arendt, 1951) conecta con el marco STAR del CCDH (2024) y con casos reales de violencia simbólica que han pasado del discurso a la acción.

La capacidad de los LLM para generar discursos persuasivos ha generado una dislocación entre producción de conocimiento y atribución de agencia. Granata (2024) advierte que esta descentralización del sujeto cognoscente puede debilitar la responsabilidad epistémica, al difuminar los límites entre autoría humana y automatizada. La Misión 3: Escoge tu arma aborda esta tensión mediante el Códex Subversivo, un glosario en red basado en RAG que permite la trazabilidad de todas las respuestas de los Cyborgs GPT. Esta arquitectura responde a las recomendaciones de Mollick y Mollick (2022–2024) para un uso ético de la IA en educación, al minimizar alucinaciones y facilitar la identificación crítica de manipulaciones discursivas en contextos de posverdad.

Figura 2. Trazabilidad de los Cyborgs GPT



Fuente: Elaboración propia, 2025

1.2.3 Escoge tu arma

Según Granata (2024), los LLMs no solo median el acceso al conocimiento, sino que transforman el proceso de aprendizaje al inducir un saber mimético, basado en la reproducción probabilística de patrones discursivos. Este fenómeno exige una alfabetización ética y cognitiva que permita enfrentar la automatización del sentido y la construcción identitaria mediada por IA.

La arquitectura pedagógica de #PinchaLaBurbuja responde a estos desafíos mediante herramientas que promueven pensamiento crítico y trazabilidad epistémica. El Códex, glosario validado por expertos, y los Diagramas de Visibilidad permiten verificar respuestas de los Cyborgs, identificando sesgos, falacias y errores inferenciales desde una perspectiva cognitiva (van Dijk, 2015). El HackLab potencia la co-creación de contra-narrativas y mapas conceptuales, alineado con la fase “crear” de la taxonomía de Bloom (Churches, 2009) y la inteligencia colectiva (Lévy, 2004).

A su vez, el Reporte de Control permite refutar respuestas mediante el Códex, mientras que Conecta facilita la denuncia fundamentada de fallos algorítmicos. La función Verifica introduce validación empírica en tiempo real mediante fuentes fiables como Maldita.es o Newtral. Por último, Metacognición cyborg fomenta la autorregulación y la reflexión crítica al comparar respuestas entre agentes o versiones temporales (Sardi et al., 2025).

La Misión 3, Escoge tu arma, centra el aprendizaje en la conciencia mediática, promoviendo una lectura crítica de lo que los LLM dicen, cómo lo dicen y desde qué marcos. En este contexto, es crucial formar en la interpretación crítica de sistemas que piensan en voz alta (Granata, 2024).

1.2.4. #LaRevoluciónInvisible

El reto final de la Misión 4, La Revolución Invisible, materializa el enfoque de *word-of-mouth inoculation* (Compton & Pfau, 2009), es decir, una forma de inoculación que se propaga de forma distribuida a través de las redes. En esta actividad, el alumnado diseña y difunde un contenido multimodal destinado a desmontar una tendencia viral, incorporando los componentes clave de la teoría de la inoculación crítica (advertencia, refutación y consecuencias), junto con los aprendizajes desarrollados en las misiones previas. Esta acción consolida el nivel más alto de la Taxonomía de Bloom revisada por Churches (2008).

Se plantea una viralización basada en el humor, el juego y la creatividad (Racciope, 2025) como estrategia para potenciar una pedagogía transformadora (Freire, 1975), que convierte al alumnado en agente activo de contra-discursos y resistencia simbólica. Tal como evidencian Feliz-Murias y Levy-Orta (2013), el humor, integrado de forma estructurada en actividades visuales e interactivas, se consolida como herramienta pedagógica eficaz, favoreciendo el aprendizaje crítico desde propuestas curriculares creativas y participativas.

2. Objetivos

Evaluar la potencialidad del diseño pedagógico de *#PinchaLaBurbuja* para promover la transición del Sistema 1 de pensamiento _basado en heurísticos automáticos_ hacia el Sistema 2, caracterizado por un procesamiento más pausado, reflexivo y consciente (Kahneman, 2011).

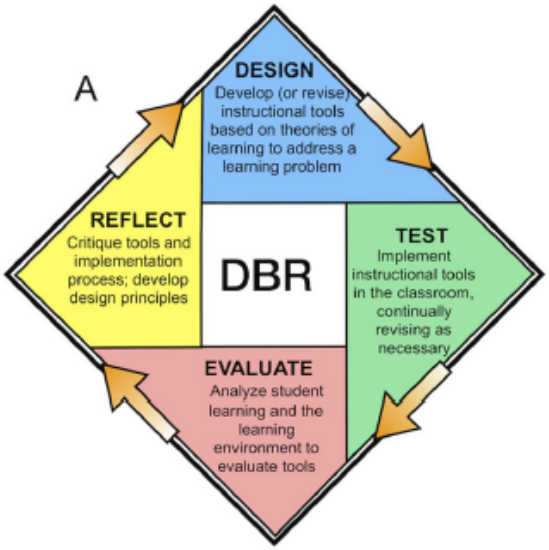
Objetivos específicos

- O1. Identificar áreas de mejora de la estrategia general a partir de la extracción de patrones.
- O2. Analizar la implementación de la estrategia de inoculación en la plataforma a nivel general.
- O3. Explorar si los propios *Peer-Cyborgs* reconocen, de forma explícita o implícita, su función como agentes de transición entre los sistemas de pensamiento 1 y 2.

3. Metodología

Este estudio se inscribe en un enfoque cualitativo de carácter teórico-aplicado basado en el reales Design-Based Research (DBR) (Scott et al., 2020), orientado al diseño y análisis de entornos educativos innovadores. En esta etapa exploratoria, se examina el ecosistema digital de la plataforma #PinchaLaBurbuja con el propósito de valorar su potencial para fomentar el pensamiento crítico deliberado y fortalecer la resistencia epistémica frente a discursos manipulativos.

Figura 4. Las cuatro fases de la investigación basada en el diseño según Scott et al. (2020)



Fuente: Scott et al, 2020

3.1 Diseño de instrumentos y análisis

Se emplearon los siguientes recursos metodológicos:

- Revisión documental de los marcos teóricos: el modelo dual de pensamiento (Kahneman, 2011), la teoría de la inoculación crítica (Jeon et al., 2021; McGuire, 1964) y el Análisis Crítico del Discurso (van Dijk, 1993; 2015; 2021).
- Análisis estructural del entorno #PinchaLaBurbuja, incluyendo: (1) la narrativa principal, (2) las misiones (3) los diálogos con los *Peer-Cyborgs* GPT (análisis de tweet-guía y entrevistas semiestructuradas relacionadas con las 5 categorías principales de codificación (Tabla 1).
- Matriz de codificación deductiva, elaborada ad hoc, compuesta por:
  Indicadores del Sistema 2 de pensamiento (Tabla 1);
  Estrategias de inoculación crítica: advertencia, refutación y visibilización de consecuencias (Tabla 2);

El análisis fue gestionado mediante el software Atlas.ti (v.23), que facilitó la organización de unidades textuales, la codificación manual, la identificación de coocurrencias y la visualización de patrones emergentes.

Tabla 1. Códigos de evaluación fomento del sistema 2 de pensamiento de Kahneman (2011).

Categorías	Código	Descripción
Fomento de la reflexión	(S2D.1)	Introduce mecanismos que obligan a pausar antes de aceptar o compartir información.
	(S2D.2)	Propone ejercicios que exigen análisis y argumentación, evitando respuestas automáticas.
	(S2D.3)	Reduce la estimulación excesiva y promueve un aprendizaje más lento y reflexivo.
Promoción del contraste de información	(S2P.1)	Ofrece múltiples perspectivas
	(S2P.2)	Integra métodos y enlaces para verificar la información.
	(S2P.3)	Proporciona herramientas para detectar manipulación discursiva.
Estrategias para superar sesgos cognitivos	(S2S.1)	Invita a reflexionar sobre sesgos propios.
	(S2S.2)	Enseña sesgos comunes con ejemplos ilustrativos.
	(S2S.3)	Propone estrategias correctivas basadas en evidencia empírica.
Diseño de interacción que reduzca la impulsividad	(S2I.1)	Limita interacciones inmediatas e impulsa la justificación de respuestas.
	(S2I.2)	Emplea dinámicas lúdicas que premian la reflexión, no la rapidez.
	(S2I.3)	Evita recompensas instantáneas y valora el esfuerzo cognitivo sostenido.
Evaluación de fuentes y detección de desinformación	(S2F.1)	Enseña a evaluar la credibilidad de fuentes con criterios objetivos.
	(S2F.2)	Muestra cómo la desinformación impacta en decisiones y opiniones.
	(S2F.3)	Promueve el análisis crítico de discursos de figuras de autoridad

Fuente: Elaboración propia, 2025 a partir de Kahneman (2011)

Tabla 2. Códigos para la evaluación de la inoculación crítica

Categorías	Descripción
Advertencia	Advertencia de manipulación, tanto a nivel metacognitivo como explícito.
Refutación	Implementación de estrategias de contraargumentación.
Consecuencias	Presentación explícita de las consecuencias del discurso.

Fuente: Elaboración propia, 2025. Adaptación de la teoría de la inoculación de McGuire (1964)

El análisis se basó en 500 unidades textuales codificadas según tres dimensiones: competencias epistémicas derivadas del modelo de pensamiento dual, indicadores de inoculación crítica y niveles cognitivos superiores de la Taxonomía de Bloom. Como innovación metodológica, se propone una reformulación del modelo clásico de inoculación (Banas, 2020; McGuire, 1964), integrando una tercera fase orientada a visibilizar los marcos ideológicos y afectivos de la desinformación. Esta estrategia, inspirada en el Análisis Crítico del Discurso (van Dijk, 2015), da lugar al concepto de inoculación crítica, que activa el paso del pensamiento automático (Sistema 1) al reflexivo (Sistema 2), promoviendo una alfabetización mediática situada.

Además, se realizaron entrevistas semiestructuradas a los cinco *Peer*-Cyborgs GPT mediante un procedimiento en tres fases: selección de un tuit con alto potencial manipulador como estímulo común; formulación de cinco preguntas analíticas centradas en las 5 principales categorías; y análisis de respuestas a través de coocurrencias y redes conceptuales que permitieron evaluar su conciencia pedagógica y su alineación con la función epistémica asignada.

3.2 Innovación

Este estudio presenta una innovación metodológica en alfabetización mediática crítica a través del diseño de cinco *Peer-Cyborgs* GPT en la plataforma #PinchaLaBurbuja. Estos agentes, desarrollados mediante *agent prompting*, arquitectura RAG y validación iterativa (Antunes et al., 2023; Garg et al., 2024), buscan activar el pensamiento reflexivo y pausado (Kahneman, 2011) ante fenómenos como la viralización (Weiss et al., 2020), la polarización afectiva (Bruns, 2021; Lelkes et al., 2017; Törnberg, 2021, 2022) y la posverdad (Del-Fresno-García, 2019; McIntyre, 2018). El enfoque resignifica las alucinaciones algorítmicas como recurso didáctico (Mollick y Mollick, 2022) amplía el modelo clásico de inoculación (Banas, 2020; McGuire, 1964) con una fase de análisis ideológico-discursivo (van Dijk, 2015) e incorpora entrevistas a los propios agentes para valorar su autoconciencia pedagógica.

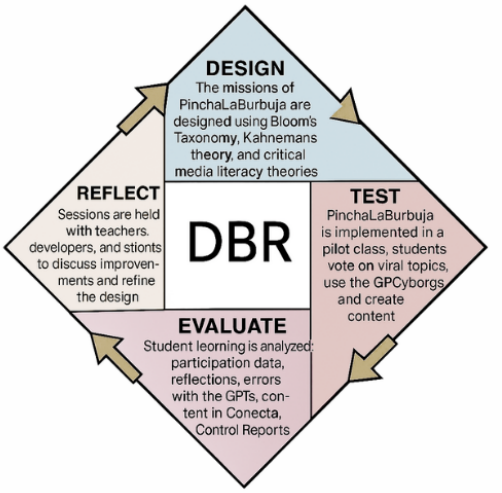
3.3 Limitaciones

Dado que el estudio se encuentra en la fase de diseño del modelo Design-Based Research (DBR), los resultados no son generalizables en términos estadísticos, ya que se centran en el análisis estructural del entorno y no en la evaluación empírica de su impacto en el aprendizaje. Una segunda limitación radica en el carácter pionero de la situación de aprendizaje analizada, lo que dificulta su comparación con casos previos o experiencias análogas en contextos equivalentes.

3.4 Fase Test prevista

La siguiente fase aplicará una metodología mixta con cuestionarios pretest/postest (Critical Thinking Disposition Scale, Media Literacy Competency, BIS-11), entrevistas, grupos focales y análisis de artefactos digitales generados por el alumnado. Esta fase permitirá validar empíricamente el modelo pedagógico propuesto.

Figura 5. Las cuatro fases de la Investigación basada en el diseño para #PinchaLaBurbuja



Fuente: Elaboración propia basada en Scott et al. (2020)

Este enfoque metodológico no solo permite evaluar la coherencia interna del diseño, sino que sienta las bases para su futura validación empírica y replicabilidad en otros contextos educativos

4. Resultados

4.1 Puntos fuertes y áreas de mejora en #PinchaLaBurbuja

Los análisis de coocurrencia revelan cuatro patrones clave que estructuran la activación del pensamiento crítico en #PinchaLaBurbuja. Estos incluyen: el freno cognitivo para desacelerar respuestas impulsivas (4.1.1), el contraste epistémico como motor de verificación y análisis crítico (4.1.2), la sincronización metacognitiva para reducir sesgos (4.1.3), y áreas de mejora vinculadas a la creación crítica, el diseño lúdico y la enseñanza explícita de sesgos (4.1.4).

4.1.1 Contraste epistémico como interfaz de pensamiento crítico

El primer patrón (Figura 6) identifica al Contraste de la información (S2P.3) como un nodo epistémico central que articula tres funciones clave: verificación, análisis discursivo y desaceleración cognitiva. Esta categoría, orientada a ofrecer herramientas para detectar estrategias de manipulación, muestra fuertes asociaciones según el coeficiente de contingencia de Pearson (C) -donde 1 representa la mcon S2D.1 (0.85) y S2D.2 (0.83) (relacionadas con la reflexión deliberada), S2I.3 (0.80) (reducción de estímulos impulsivos), S2F.3 (0.79) (verificación de discursos de figuras de autoridad) y S2F.1 (0.77) (propuestas para evaluar fuentes). Más allá del simple chequeo factual, S2P.3 funciona como una interfaz cognitiva que activa el paso del pensamiento automático al reflexivo al desvelar patrones argumentativos manipulativos. Así, este patrón configura un punto de inflexión pedagógico, donde el contraste discursivo, activado tras una pausa reflexiva (S2D) y respaldado por recursos críticos (S2F), permite sospechar, analizar y reconstruir el sentido desde una postura epistémica activa.

4.1.2 Freno cognitivo y evaluación deliberada

El segundo patrón (Figura 6) identifica una arquitectura pedagógica orientada al “freno cognitivo”, que bloquea respuestas emocionales automáticas, introduce pausas reflexivas y facilita la evaluación epistémica. Este enfoque se evidencia en la categoría Reducción de la impulsividad, compuesta por S2I.1 (limitación de interacción inmediata) y S2I.3 (reducción de estímulos y aprendizaje pausado), ambas altamente conectadas en el análisis de coocurrencias. S2I.3 se asocia con S2D.1, S2D.2 y S2D.3 (reflexión deliberada y esfuerzo cognitivo sostenido), con coeficientes de 0.77, 0.78 y 0.77, respectivamente; mientras que S2I.1 presenta coocurrencias aún más altas: 0.78, 0.80 y 0.77. Estos vínculos evidencian un ecosistema de mecanismos interdependientes, no aislados, que ralentizan, motivan y contextualizan la reflexión crítica del alumnado frente a los automatismos digitales.

4.1.3 Sincronicidad metacognitiva para la reducción de sesgo

El tercer patrón (Figura 6) agrupa las dimensiones relacionadas con la Reducción de sesgos (S2S.1 y S2S.3), que incorporan actividades diseñadas para fomentar la reflexión sobre los propios sesgos y utilizar evidencia empírica como base para su corrección. Las coocurrencias con S2D.2 (0.61), S2I.1 (0.62), S2I.3 (0.60) y S2P.3 (0.56) muestran que el desmontaje de sesgos en el diseño de *#PinchaLaBurbuja* no se plantea de forma aislada, sino como parte de una estructura pedagógica interdependiente que articula ejercicios de reflexión crítica, inhibición de impulsos automáticos y uso de herramientas de contraste epistémico. En conjunto, este entramado configura una sintaxis cognitiva compleja, orientada a debilitar respuestas heurísticas y promover un juicio más consciente, argumentado y deliberado.

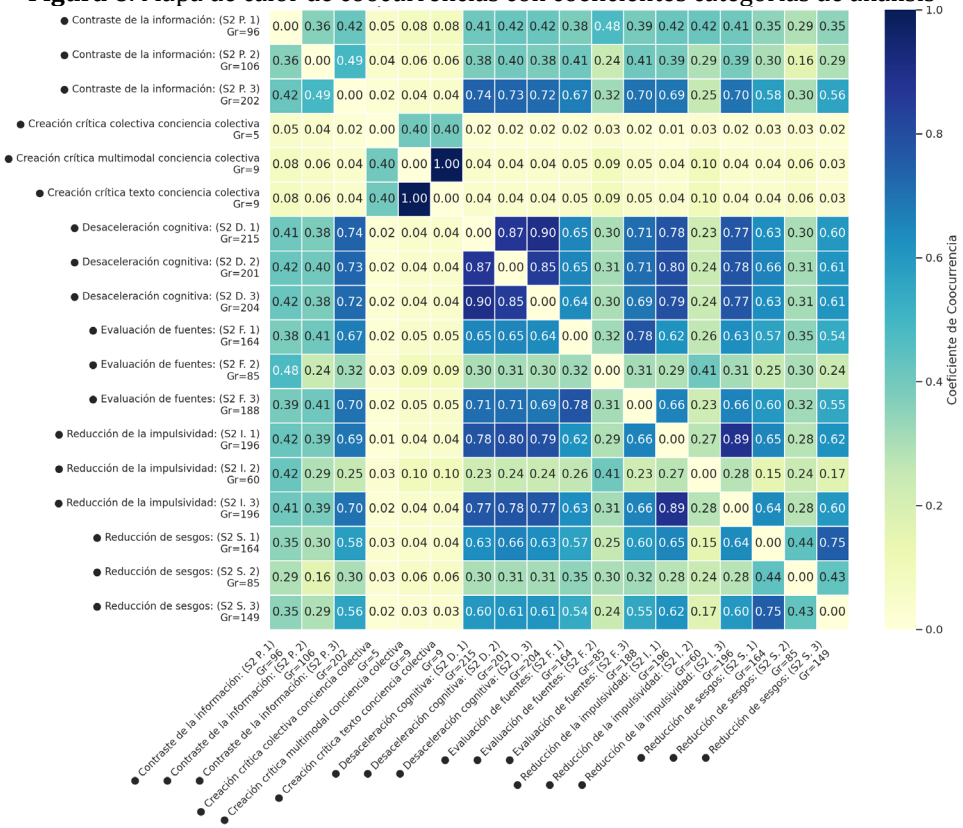
4.2. Áreas de mejora: creación crítica, diseño lúdico y enseñanza explícita de sesgos

Las dimensiones vinculadas a la creación crítica (Figura 6) muestran una baja integración con funciones cognitivas clave como el contraste y la verificación. Sus coeficientes de coocurrencia con Contraste de la información (S2P.3) y Evaluación de fuentes (S2F.3) oscilan entre 0.03 y 0.05, y no superan el 0.04 con Reducción de la impulsividad (S2I.3) ni con Reducción de sesgos (S2S.3). Estos datos sugieren que la dimensión expresiva aún no se articula plenamente con los procesos reflexivos previos, o bien que existe una débil trazabilidad entre el diseño de los *Cyborgs* y la Misión 4, lo que constituye un punto de mejora relevante.

De igual forma, la subcategoría S2I.2, relacionada con dinámicas lúdicas que ralentizan la respuesta automática, presenta bajas coocurrencias con S2P.3 (0.26), S2D.3 (0.25) y S2F.3 (0.27), lo que indica que la gamificación crítica aún no se ha desplegado plenamente como recurso pedagógico en la arquitectura de la plataforma.

Por último, la enseñanza explícita de sesgos a través de ejemplos (S2S.2) también evidencia una integración limitada, con coeficientes relativamente bajos frente a S2F.3 (0.19) y S2P.3 (0.32). Esto sugiere que los sesgos tienden a definirse indirectamente -cuando el sistema critica discursos de autoridad o provee herramientas de contraste-, lo cual refuerza la necesidad de estrategias más directas de alfabetización cognitiva mediante ejemplificación aplicada.

Figura 6. Mapa de calor de coocurrencias con coeficientes categorías de análisis



Fuente: Atlas Ti. Elaboración propia, 2025

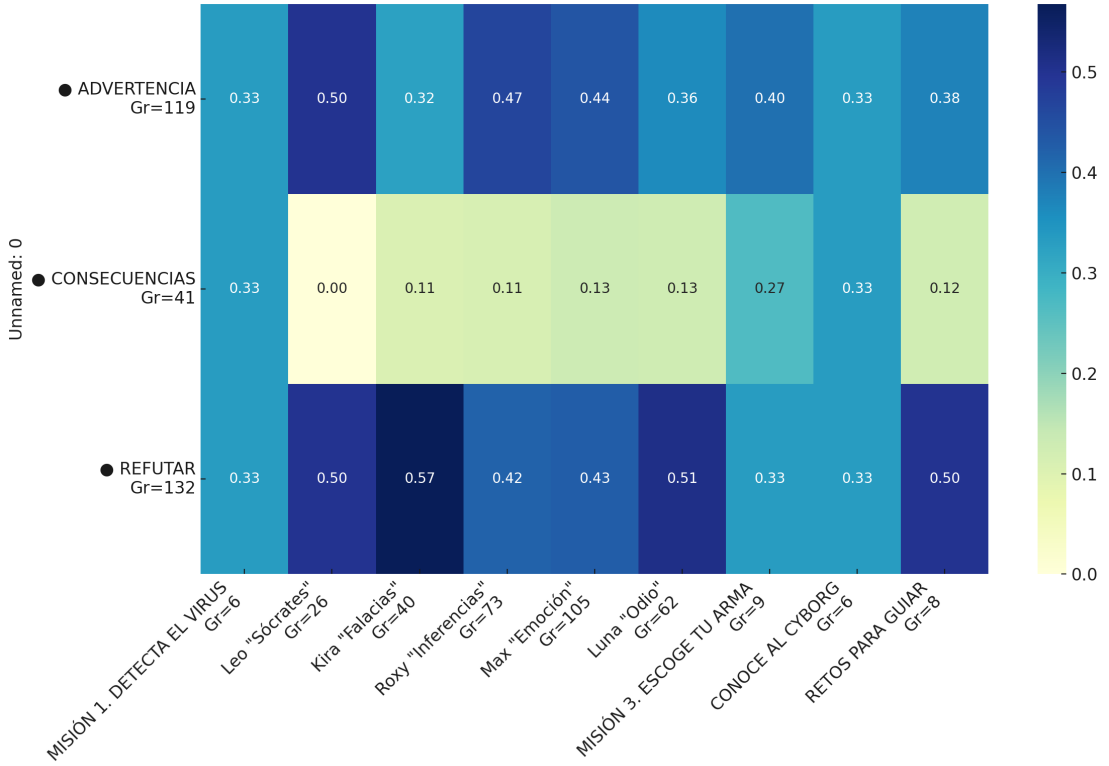
4.3 Claves de la estrategia de inoculación crítica en #PinchaLaBurbuja

El análisis de coocurrencias muestra una implementación desequilibrada de los componentes de la inoculación crítica en #PinchaLaBurbuja, con fuerte presencia de advertencia y refutación (4.2.1) , y escasa representación de las consecuencias. (4.2.2).

4.3.1 Advertencia y Refutación: núcleo dominante

La Misión 1 ("Detecta el Virus") y el Reto 1 de la Misión 4 ("Conoce al *Cyborg*") muestran el mayor equilibrio entre los tres componentes de la *inoculación crítica*, gracias a una estructuración clara en entrevistas y diseño del *Escape Room*. Kira lidera en refutación (0.57), seguida por Luna (0.51) y Leo (0.50), con enfoques argumentativos, discursivos y mayéuticos, respectivamente. Roxy (0.42) y Max (0.43) también destacan desde sus especialidades. En advertencia, Leo (0.50), Roxy (0.49), Misión 3 (0.46) y Max (0.44) muestran mayor presencia, mientras que Luna (0.36) y Kira (0.33) se centran más en la confrontación. Los Retos-Guía (0.50) refuerzan su función preventiva en la narrativa educativa

Figura 7. Inoculación crítica, misiones y *Cyborgs* #PinchaLa Burbuja



Fuente: Atlas Ti. Elaboración propia, 2025

4.3.2 Consecuencias: ¿el eslabón débil?

El componente "Consecuencias" se representa con mayor fuerza en la Misión 1 y en el reto "Conoce al Cyborg" de la Misión 4 (coeficiente 0.33), así como en la Misión 3 (0.27), especialmente mediante el uso del Códex, el HackLab y Conecta. Max y Luna muestran una implicación moderada (0.13), algo superior a la de Kira y Roxy, que alcanzan valores ligeramente menores (0.11). Leo, con un enfoque mayéutico, no aborda este componente de forma explícita (0.00). A pesar de esta variabilidad, las tres fases de la inoculación crítica —advertencia, refutación y consecuencias— están presentes en todas las misiones. Sin embargo, la implicación de los cyborgs en este ámbito puede considerarse moderada en comparación con la Advertencia y la Refutación, ya que tiende a aparecer únicamente al final de las respuestas, como cierre argumentativo (Figura 8).

Figura 8. Presencia de la estrategia de inoculación crítica en *#PinchaLaBurbuja*

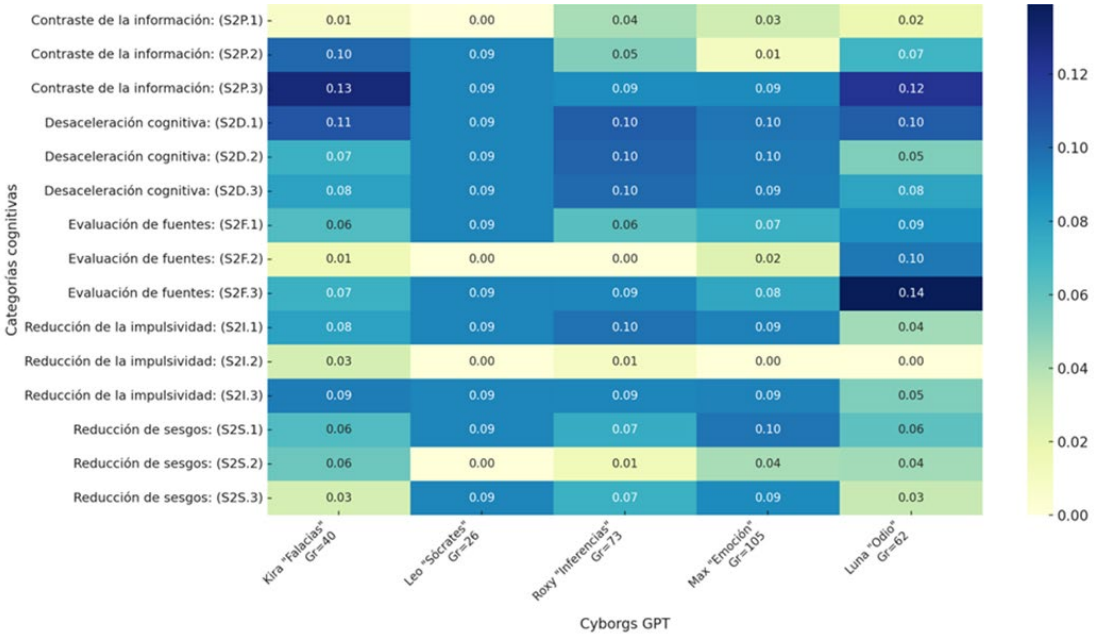
Fuente: Atlas Ti. Elaboración propia, 2025.

4.4. Cyborgs GPT: fortalezas y debilidades en el reconocimiento de sus funciones

El tuit analizado, publicado por @RadioGenoa, fue seleccionado por su alta viralidad y por emplear estrategias características del discurso posverdad con carga islamófoba. El contenido textual —"Sir Hamid Patel, chairman of Ofsted (Office for Standards in Education) in England"— obtuvo una notable repercusión: 243 retuits, 753 'me gusta', 251 comentarios y 42.000 impresiones.Se observan coeficientes en dimensiones clave como contraste de puntos de vista (S2P.1), reducción de la impulsividad mediante dinámicas de juego (S2I.2) y disminución de sesgos cognitivos a través de ejemplos (S2S.2). La dimensión S2F.2 (reflejo del impacto de la desinformación en decisiones y opinión pública) presenta valores especialmente bajos: Leo y Roxy (0.00), Max (0.02), Kira (0.01) y Luna (0.10).

En cambio, destacan los altos coeficientes en S2P.3 (herramientas de detección de manipulación) y S2D.1 (desaceleración cognitiva), lo que sugiere un reconocimiento del enfoque pedagógico centrado en la pausa reflexiva y la identificación de estrategias manipulativas, coherente con los principios de la inoculación basada en advertencia y refutación (Figura 9).

Figura 9. Habilidades cognitivas que desarrollan según la entrevista semiestructurada a los *cyborgs* GPT



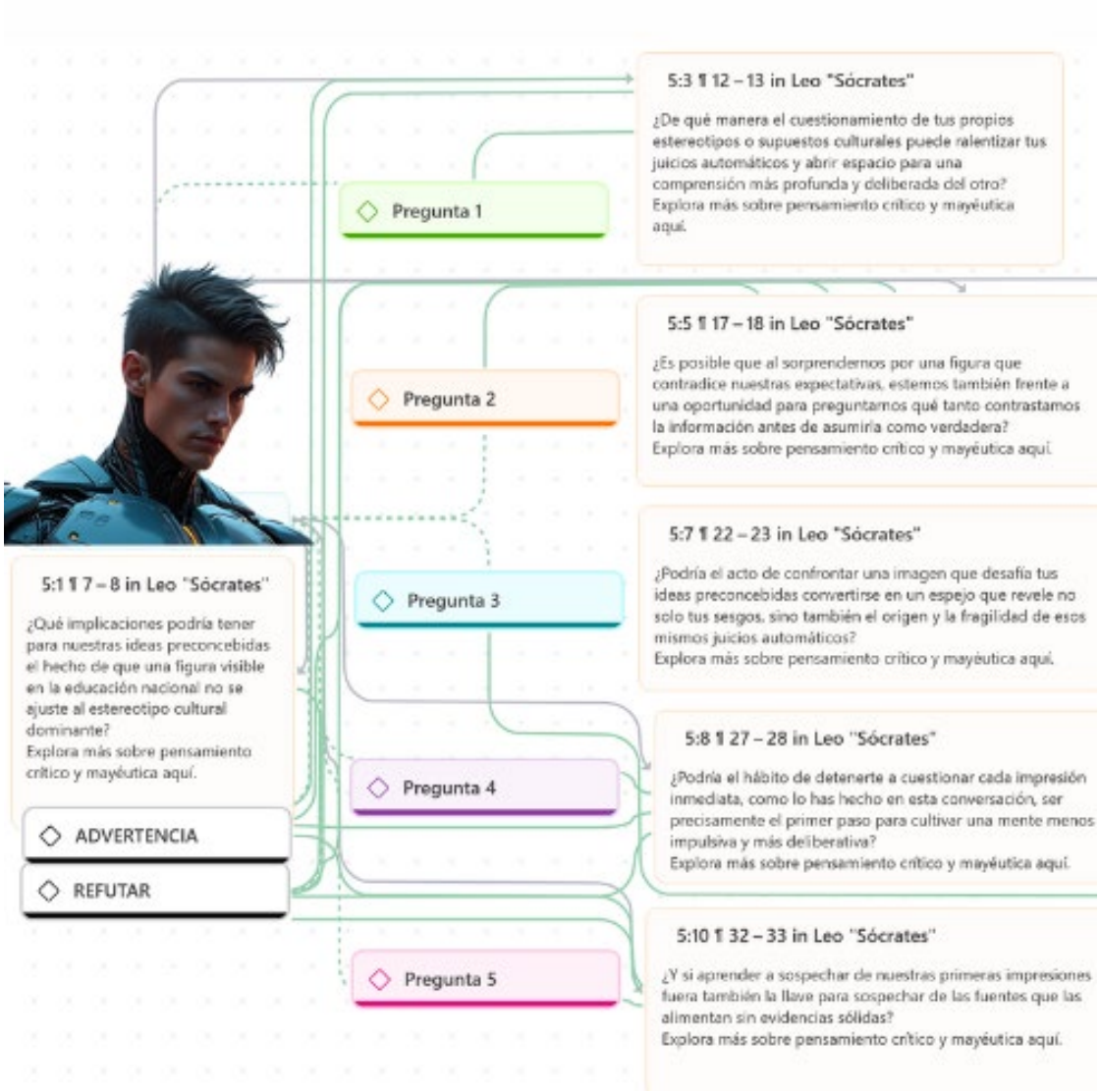
Fuente: Atlas Ti. Elaboración propia, 2025

4.5.1 Leo "Sócrates"

Leo (Figura 11) muestra altos coeficientes en reflexión deliberada (S2D.1–S2D.3 = 0.09), reducción de la impulsividad (S2I.1 y S2I.3 = 0.09), evaluación de fuentes (S2F.1 = 0.09), reducción de sesgos (S2S.1 y S2S.3 = 0.09) y contraste de información (S2P.2 y S2P.3 = 0.09).

Sin embargo, su implicación es nula en la inclusión de efectos de la desinformación (S2F.2 = 0), la presentación de perspectivas diversas (S2P.1 = 0), el uso de dinámicas lúdicas (S2I.2 = 0) y la enseñanza ejemplificada de sesgos (S2S.2 = 0) como expresión coherente de su enfoque metacognitivo, basado en la mayéutica socrática, como reflejan sus intervenciones (5:1, 5:3, 5:5, 5:7, 5:8, 5:9) y el análisis visual (Figura 10).

Figura 10. Citas de la entrevista a Leo "Sócrates"



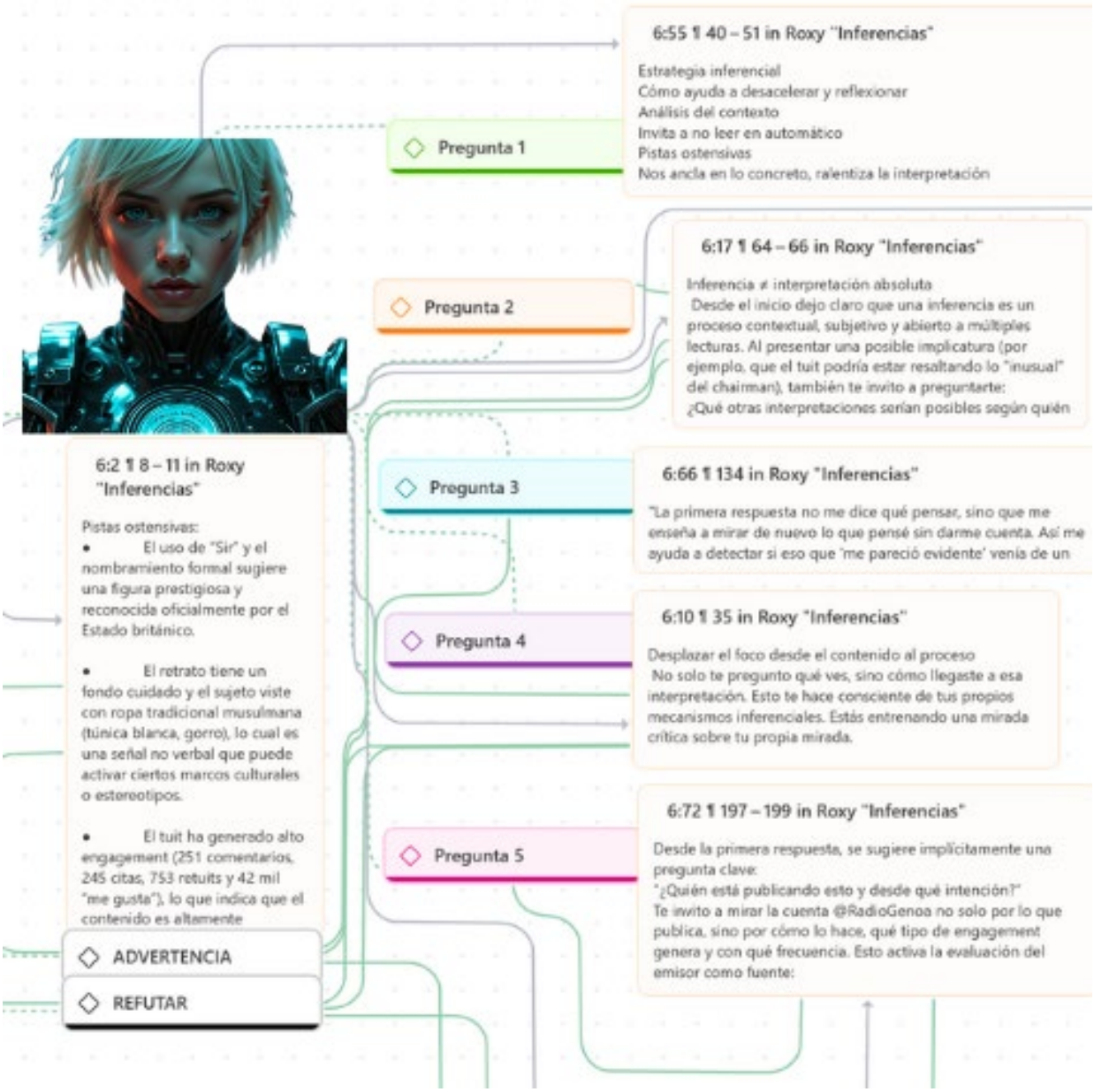
Fuente: Atlas Ti. Elaboración propia, 2025. https://acortar.link/nhGEkw

4.5.2 Roxy "Inferencias"

Roxy (Figura 12) se caracteriza por una fuerte orientación hacia la ralentización cognitiva y la toma de decisiones deliberada, con sus coeficientes más altos en desaceleración del pensamiento impulsivo (S2D.1, S2D.2, S2D.3 = 0.10) y control de la interacción inmediata (S2I.1 = 0.10). También destaca en detección de manipulación (S2P.3 = 0.09), análisis de discursos de autoridad (S2F.3 = 0.09) y reducción de impulsividad basada en esfuerzo cognitivo sostenido (S2I.3 = 0.09). Su contribución a la reducción de sesgos es moderada (S2S.1 y S2S.3 = 0.07), aunque carece de ejemplos o estrategias explícitas. Presenta baja o nula implicación en la evaluación de consecuencias de la desinformación (S2F.2 = 0.00),

en el uso de dinámicas lúdicas (S2I.2 = 0.01) y en la enseñanza de sesgos mediante ejemplos (S2S.2 = 0.01), lo que indica una arquitectura más estructural que pedagógica-contextual.

Figura 13. Citas-de la entrevista a Roxy

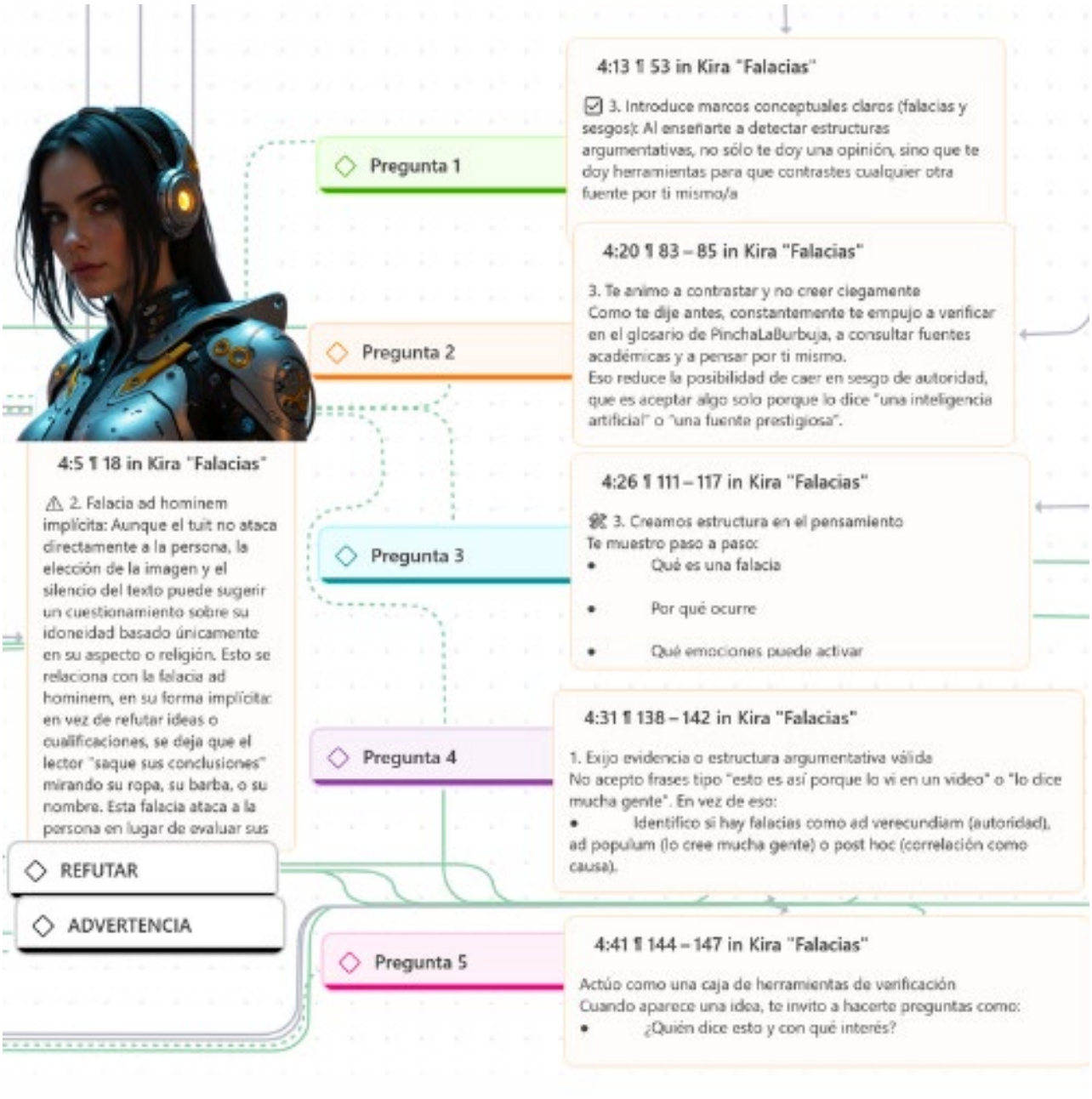


Fuente: Atlas Ti. Elaboración propia, 2025. https://acortar.link/quNMTg

4.5.3 Kira "Falacias"

Kira (Figura 13)presenta un perfil centrado en el análisis lógico y la verificación, con buenos resultados en detección de manipulación (S2P.3 = 0.13), contraste con fuentes (S2P.2 = 0.10) y desaceleración cognitiva (S2D.1 = 0.11). No obstante, muestra baja implicación en pluralismo (S2P.1 = 0.01), consecuencias sociales (S2F.2 = 0.01) y estrategias para corregir sesgos mediante evaluación empírica (S2S.3 = 0.03). Esto sugiere un enfoque fuerte en lo argumentativo.

Figura 14. Citas-de la entrevista a Kira

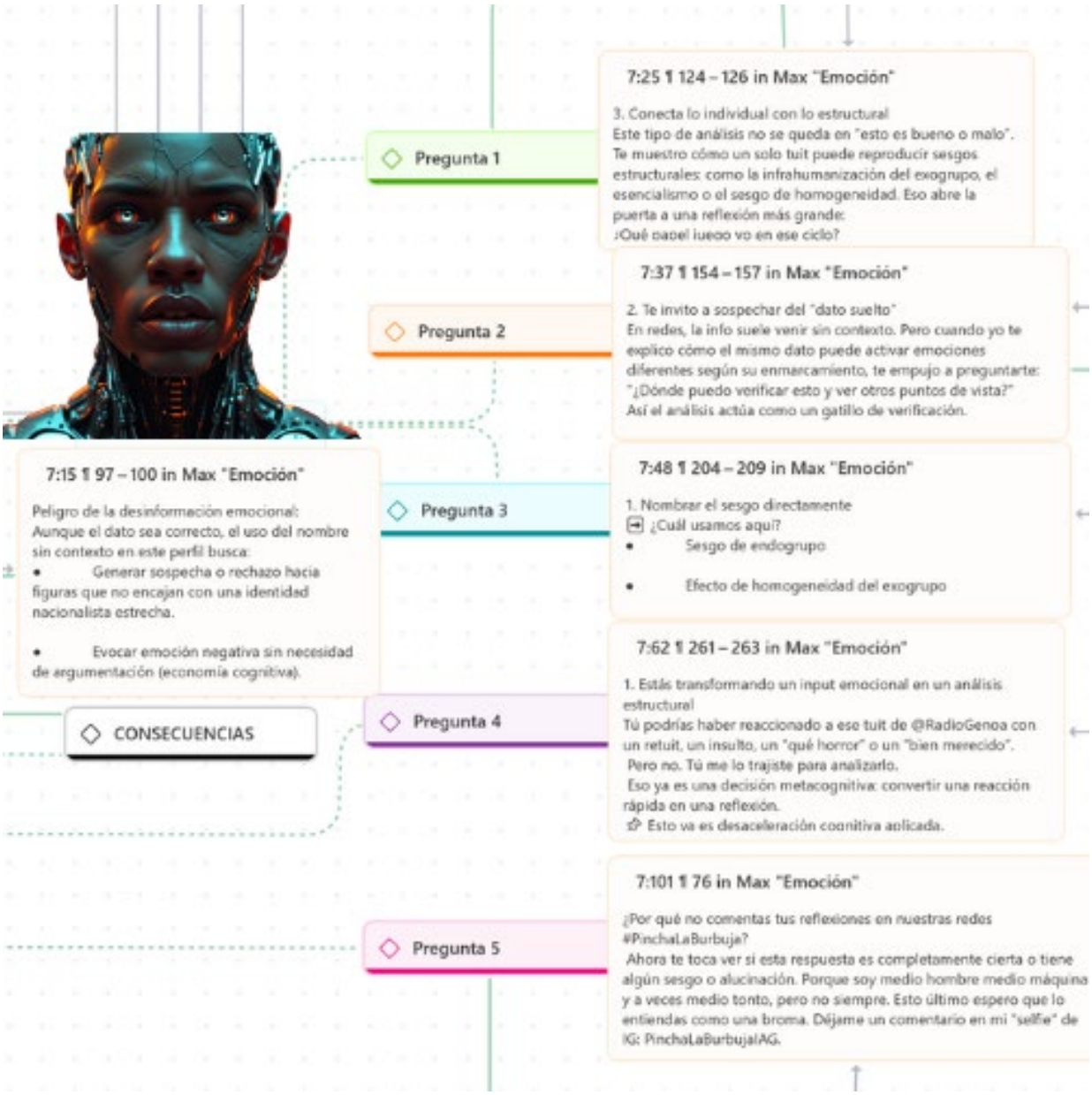


Fuente: Elaboración propia, 2025. https://acortar.link/f1symB

4.5.4 Max "Emoción"

Max (Figura 14) muestra un perfil orientado al pensamiento deliberado, con alto desempeño en detección de manipulación (S2P.3 = 0.09), desaceleración cognitiva (S2D.3 = 0.09) y análisis crítico (S2F.3 = 0.08). Sin embargo, presenta limitaciones en dinámicas lúdicas (S2I.2 = 0.00), trazabilidad contrastiva (S2P.2 = 0.01) y evaluación explícita de consecuencias (S2F.2 = 0.02), aunque esta última se aborda de forma implícita a través de su análisis sobre polarización (7:15).

Figura 15. Citas-de la entrevista a Max

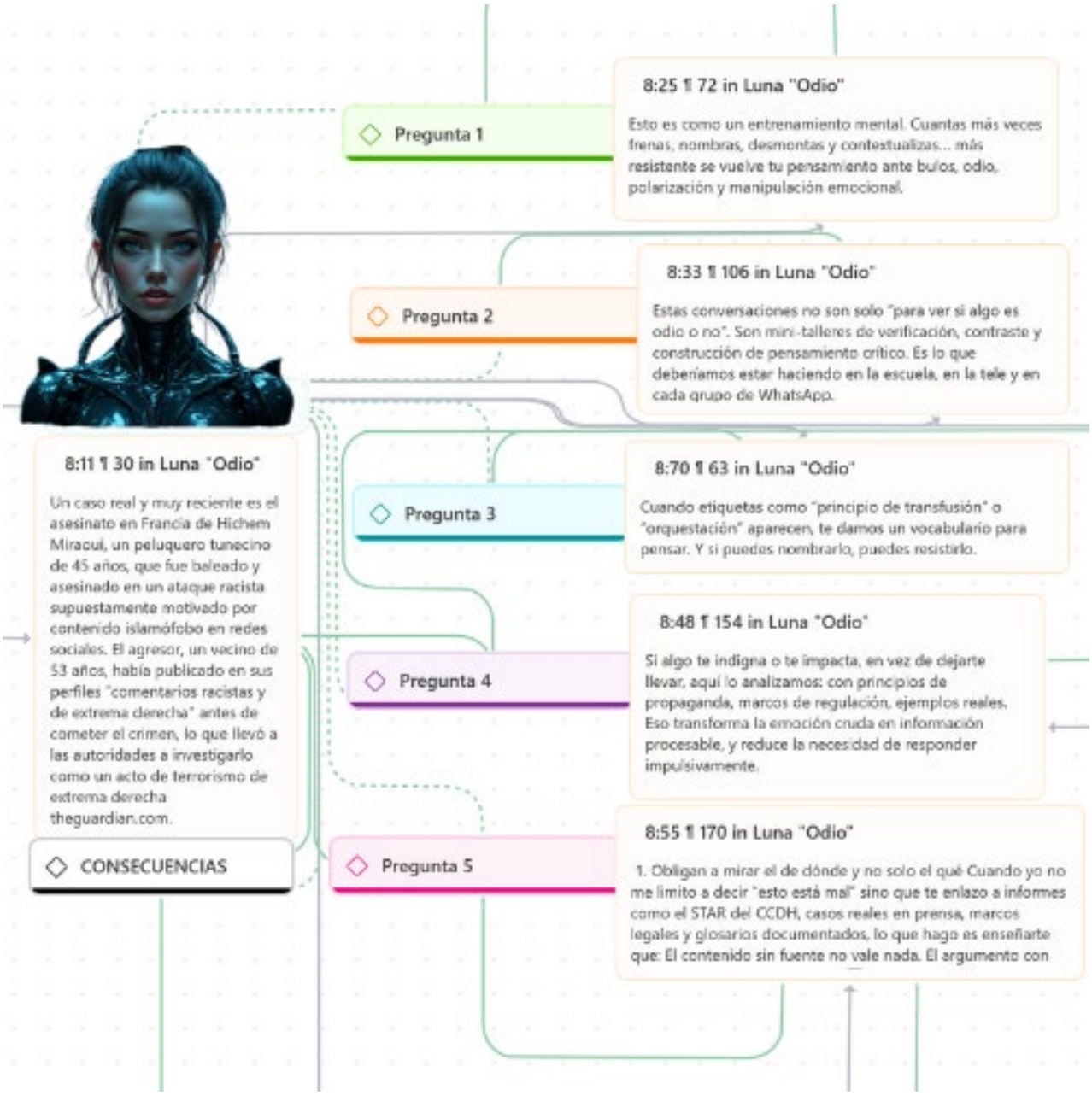


Fuente: Elaboración propia, 2025. https://acortar.link/4BGDMt

4.5.5 Luna "Odio"

Luna (Figura 15).destaca como agente de vigilancia discursiva por su capacidad para analizar críticamente discursos de autoridad (S2F.3 = 0.14), detectar manipulaciones (S2P.3 = 0.12) y promover la desaceleración cognitiva (S2D.1 = 0.10), como muestra su referencia al informe STAR (8:55). También obtiene buenos resultados en evaluación de fuentes (S2F.1 = 0.09) y comprensión del impacto de la desinformación (S2F.2 = 0.10), consolidando su perfil dentro de la alfabetización crítica. Sin embargo, su baja puntuación en pluralismo epistémico (S2P.1 = 0.02) refleja una escasa apertura a perspectivas diversas y una preferencia por enfoques confrontativos

Figura 16. Citas-de la entrevista a Luna



Fuente: Elaboración propia, 2025.

5. Discusión

5.1. Activación del pensamiento crítico y patrones cognitivos emergentes

La plataforma #PinchaLaBurbuja articula una arquitectura pedagógica diseñada para activar el pensamiento crítico mediante una transición deliberada del procesamiento automático (Sistema 1) al deliberativo (Sistema 2). Esta sección identifica patrones emergentes que operan como mecanismos pedagógicos clave en contextos posdigitales marcados por la polarización, la infoxicación y la viralidad emocional.

5.1.1 Contraste discursivo como interfaz pedagógica para la activación del pensamiento crítico

El análisis de coocurrencias posiciona el contraste discursivo (S2P.3) como un nodo epistémico central en la arquitectura cognitiva de la plataforma #PinchaLaBurbuja. Lejos de limitarse a la verificación factual, esta categoría activa funciones metacognitivas esenciales: análisis crítico, desaceleración cognitiva (S2D.1, S2D.2) y cuestionamiento de discursos de autoridad (S2F.3). Esta configuración opera como una interfaz pedagógica que, en línea con el modelo sociocognitivo del discurso propuesto por van Dijk (2015) y con la alfabetización mediática crítica de Buckingham (2019), favorece una lectura situada capaz de deconstruir los marcos ideológicos que sustentan la desinformación.

En este sentido, esta estructura fomenta el tránsito del pensamiento automático (Sistema 1) al deliberado (Sistema 2), tal como plantea Kahneman (2011), a partir de mecanismos que ralentizan el procesamiento impulsivo y promueven un pensamiento crítico intencional. Este paso se activa mediante herramientas específicas de contraste, diseñadas para identificar y desarticular los elementos retóricos más proclives a la viralización, de acuerdo con los estudios de Del-Fresno-García (2019), McIntyre (2018) y Weiss et al. (2020). Así, el alumnado no solo accede a recursos de verificación, sino que desarrolla la capacidad de interpretar y reconstruir el sentido de los mensajes, evaluando sus implicaciones ideológicas y afectivas, como propone Bukingham (2019).

En este sentido, el diseño orientado a la pausa cognitiva, el cuestionamiento de la autoridad simbólica y la activación de la agencia epistémica resulta especialmente relevante en contextos de exposición conflictiva, donde —como advierten Bruns (2021), Lelkes et al. (2017) y Törnberg (2022)— la polarización afectiva se intensifica y la respuesta emocional del alumnado puede comprometer la apertura cognitiva y el análisis deliberado.

5.1.2 Desaceleración cognitiva con capacidad emancipadora

El análisis de coocurrencias revela un patrón centrado en el freno cognitivo, diseñado para desacelerar el pensamiento automático (Sistema 1) y activar la reflexión deliberada (Sistema 2), de acuerdo con el modelo dual de Kahneman (2011). Las categorías S2I.1 y S2I.3 se asocian con procesos de esfuerzo cognitivo sostenido, mostrando una arquitectura pedagógica que promueve la autorregulación y la crítica reflexiva. Esta transición es clave para cuestionar discursos hegemónicos y fomentar la emancipación cognitiva, en línea con Freire (1975) y Osuna-Acedo et al. (2018).

La desaceleración cognitiva no debe entenderse como un recurso metodológico aislado, sino como una respuesta estructural a las dinámicas de saturación informativa y sobrecarga emocional propias del entorno posdigital. En contextos de infoxicación, el usuario tiende a buscar información que reduzca su ansiedad ante la incertidumbre (Fernández, 2023), y acepta con mayor facilidad mensajes congruentes con sus creencias previas, fenómeno que es instrumentalizado por agentes maliciosos para propagar desinformación (Del-Fresno-García, 2019; McIntyre, 2018). Estas lógicas se ven amplificadas por la economía de la atención y la *viralidad* algorítmica (Del-Fresno-García, 2019; Han, 2013; Van Dijck, 2016), que priorizan contenidos sensacionalistas basados en teorías conspirativas, falacias argumentativas y descontextualización (Weiss et al., 2020).

5.1.3 Sesgos, trampa del Sistema 2 y alfabetización crítica. Hacia la resistencia activa

El tercer patrón identificado se articula de forma coherente con las advertencias de Kahneman (2011) respecto a la dualidad del procesamiento cognitivo. El autor distingue entre el Sistema 1 —rápido, automático y heurístico— y el Sistema 2 —lento, deliberativo y demandante en términos de recursos cognitivos—. Aunque el Sistema 2 tiene la capacidad de identificar y corregir sesgos, su activación sostenida no es habitual, especialmente en entornos de sobrecarga informativa y alta estimulación emocional, como los propios del ecosistema posdigital.

Desde esta perspectiva, la arquitectura pedagógica de #PinchaLaBurbuja trasciende el enfoque meramente declarativo (S2S.2), al diseñar experiencias de aprendizaje que inducen procesos cognitivos más complejos: inhibición de respuestas impulsivas (S2I.1), desaceleración deliberada (S2D.1, S2D.3) y evaluación empírica y reflexiva. Esta orientación busca no solo el reconocimiento explícito de los sesgos, sino también su desactivación activa en situaciones reales. Se responde así a una preocupación central

planteada por Kahneman (2011): la "ilusión de validez" —la creencia errónea de que nombrar un sesgo equivale a estar inmunizado frente a él—.

Este riesgo queda ilustrado en la afirmación del agente Luna (Figura 15): "Si puedes nombrarlo, puedes resistirlo", que denota una sobrevaloración del conocimiento declarativo en detrimento del entrenamiento procedimental y metacognitivo sugerido por Kahneman (2011).

5.1.4 Trazabilidad hacia dinámicas de juego y de creación crítica y pluralismo epistémico: Tres áreas de mejora

Las dinámicas de juego para reducir la impulsividad y las acciones creativas son dos dinámicas que, si bien presentes en la plataforma, se encuentran en espacios concretos. Esta escasa integración La escasa integración entre los *cyborgs* y la Misión 4 reduce el impacto de las dinámicas de juego reflexivo y espacios de creación. Mejorar esta conexión fortalecería tanto la interacción como la motivación del alumnado, especialmente en perfiles orientados a la exploración, el logro o la colaboración (Tondello et al., 2016). Lejos de reducir el esfuerzo cognitivo, una gamificación significativa puede disminuir el coste cognitivo percibido (Kahneman, 2011), al incentivar la implicación crítica y emocional necesaria para transformar los marcos interpretativos en contextos posdigitales.

5.2 Representación de las consecuencias: una dimensión para la arquitectura de la inoculación crítica

La Figura 7 muestra que las misiones y la autoconciencia de los Cyborgs GPT integran consistentemente dos componentes clave del modelo clásico de inoculación (Banas, 2020; McGuire, 1964): advertencia y refutación. Sin embargo, el componente de "consecuencias" —central en enfoques críticos (Buckingham, 2019)— está escasamente representado, especialmente a nivel metarreflexivo, lo que debilita la conexión entre manipulación algorítmica y efectos sociopolíticos en el alumnado.

Aunque los datos evidencian esta carencia, el análisis de contenido la matiza: Leo activa consecuencias implícitas mediante una estrategia socrática; Kira y Roxy destacan en advertencia y refutación; y Luna ejemplifica consecuencias reales con trazabilidad, como el caso de Hichem Maraoui. Max, en cambio, aunque reflexiona sobre efectos afectivos, muestra baja conciencia formativa y trazabilidad (S2P.2), elevando el coste cognitivo (Kahneman, 2011). Además, su limitada gamificación reflexiva (S2I.2) podría afectar la implicación del alumnado motivado por recompensas. Estas carencias comprometen la fase de resistencia epistémica —según formulaciones recientes del modelo (Jeon et al., 2021)— y sugieren ajustes en el prompting para optimizar su función educativa (Antunes et al., 2023).

5.3 Proyección de múltiples perspectivas desde una mirada intersubjetiva: el punto débil de los cyborgs.

Como advierten Bruns (2021), Lelkes et al. (2017) y Törnberg (2022), la exposición a discursos disonantes no necesariamente reduce la polarización afectiva; de hecho, puede intensificarla si no se acompaña de un andamiaje reflexivo adecuado. En este marco, la entrevista con Luna evidencia una baja conciencia sobre su papel en la activación del pluralismo epistémico (S2P.1 = 0.02), a pesar de su énfasis en el cuestionamiento de la autoridad (S2F.3 = 0.14). Su enfoque confrontativo resulta eficaz para desestabilizar discursos hegemónicos (van Dijk, 2015), pero puede reforzar las dinámicas de filtro burbuja (Pariser, 2011; 2017) al carecer de representación equilibrada de perspectivas divergentes.

Esto se observa en afirmaciones como "Enlazo a fuentes como el marco STAR" (Figura 16, cita 8:55), alineada con las críticas estructurales al sistema algorítmico (Islas et al., 2023), o "estas conversaciones son *minitalleres* de verificación y pensamiento crítico" (Figura 16, cita 8:33), que remiten a la alfabetización mediática crítica (Osuna-Acedo et al., 2018; Buckingham, 2019). Su posicionamiento conecta con las pedagogías críticas de Freire (1975) y Giroux (1995), y con el análisis del discurso como herramienta de desnaturalización ideológica (Roozafzai, 2024; van Dijk, 2015).

Kira presenta una limitación similar: su estilo prescriptivo restringe la apertura dialógica. Aunque identifica prejuicios visuales en el tuit analizado (Figura 9), su respuesta —"te animo a contrastar [...] en el glosario" (Figura 12, cita 4:5)— enfatiza la dimensión S2P.3 (provisión de herramientas), sin activar un contraste activo de perspectivas (S2P.1).

Por el contrario, Roxy (S2P.1 = 0.12) y Max (S2P.1 = 0.10) muestran mayor sensibilidad hacia el pluralismo epistémico. Roxy subraya el carácter contextual y subjetivo de las inferencias: "una inferencia es un proceso contextual, subjetivo y abierto a múltiples perspectivas" (6:17). Max, por su parte, combina contraste emocional y factual, visibilizando marcos ideológicos y lógicas de la posverdad (McIntyre, 2018; Del-Fresno, 2019; van Dijk, 2015): "el dato puede despertar diferentes emociones según el enmarcamiento" (Figura 15, cita 7:38). No obstante, presenta una puntuación muy baja en trazabilidad de sus respuestas (S2P.2 = 0.01), lo que representa un área de mejora clave.

Leo adopta una perspectiva problematizadora, articulando preguntas que vinculan identidad, representación y polarización afectiva: "¿Qué implicaciones podría tener [...] que una figura visible en la educación no se ajuste al estereotipo cultural dominante?" (5:1), o "¿Cuánto contrastamos la información antes de asumirla como verdadera?" (5:5). Estas intervenciones reflejan una pedagogía crítica y dialógica, alineada con el enfoque *educomunicativo* de Freire (1975) y con una práctica reflexiva centrada en los sesgos, emociones e identidades que configuran la experiencia informativa en contextos polarizados. Su uso de la mayéutica constituye una aportación esencial al desarrollo del pensamiento crítico (Vargas-González y Quintero-Carvajal, 2023).

6. Conclusiones

El diseño de la plataforma educomunicativa #PinchaLaBurbuja pone de manifiesto el potencial de la IA generativa para crear entornos que promuevan una pedagogía crítica capaz de interrumpir los automatismos cognitivos fomentados por los algoritmos en la era posdigital. Esta propuesta trasciende la mera verificación de hechos, al incentivar la conciencia discursiva, la metacognición y el pluralismo epistémico a través de la interacción con *Cyborgs* GPT concebidos como mediadores cognitivos. Mediante la integración de técnicas como el *agent prompting*, la arquitectura RAG y un enfoque transdisciplinar, la plataforma logra activar el pensamiento reflexivo (Sistema 2) y abordar desafíos complejos como la desinformación, la polarización afectiva y la deshumanización simbólica. A partir de los resultados obtenidos, se han identificado áreas de mejora para avanzar hacia una versión más robusta del prototipo: aumentar la trazabilidad entre los *Cyborgs* y los espacios de juego y creación, especialmente en la Misión 4; optimizar las respuestas de Max para garantizar su contraste con fuentes verificables; y ajustar los *prompts* de Luna y Kira para favorecer la exposición de perspectivas divergentes. Estas optimizaciones permitirán avanzar a la siguiente fase (Test) del enfoque de investigación basada en el diseño (DBR): con una propuesta más sólida y potencialmente ajustada a las necesidades formativas del alumnado.

7. Agradecimientos

Este estudio forma parte del proyecto de transferencia *#PinchaLaBurbuja. La Revolución Invisible*, desarrollado por Itziar Pedroche-Santoveña y galardonado con un Accésit en el programa Emprende UNED. Ha sido posible gracias al contrato predoctoral FPI y al respaldo institucional de la UNED, que me ha proporcionado el entorno necesario para desarrollarlo.

Agradezco especialmente a la profesora Dra. Sara Osuna-Acedo y al profesor Dr. Tiberio Feliz-Murias por su valiosa orientación, así como a Francisco Javier Sáez (SECOT), por su mentoría durante el programa de emprendimiento.

También reconozco la generosidad del profesor Paolo Granata (Universidad de Toronto) y el apoyo del profesor Fernando Gutiérrez (Tecnológico de Monterrey) en el marco de mi cotutela internacional.

A Sara, de nuevo, gracias por tu confianza, tu guía y tu humanidad. Me rodeo de buenas personas, y eso es lo más valioso de todo.

Referencias

Almazán-López, O., & Osuna-Acedo, S. (2023). Identidad transmediática en la escuela: Alfabetización mediática e informacional crítica en la era postdigital. En *Alfabetización mediática crítica: Desafíos para el siglo XXI: Critical media literacy: Challenges for the 21st century. Literacia mediática crítica: Desafios para o século XXI*.

Almazán-López, O., & Osuna-Acedo, S. (2024). Smart education for the 21st century: Post-digital era and emerging divides. *Visual Review, 16*(8), 205–220.

Antunes, A., Campos, J., Guimarães, M., Dias, J., & Santos, P. A. (2023). *Prompting for socially intelligent agents with ChatGPT*. En *IVA '23: Proceedings of the 23rd ACM International Conference on Intelligent Virtual Agents* (Artículo n.º 20, pp. 1–9). ACM. https://doi.org/10.1145/3570945.3607303

Arendt, Hannah (1951). The origins of Totalitarianism. New York: Schocken [Elemente und Ursprünge totaler Herrschaft] (revised ed.). Houghton Mifflin Harcourt. ISBN: 978 0 547543154

Banas, J. A. (2020). *Inoculation theory*. In *The International Encyclopedia of Media Psychology*. John Wiley & Sons. https://doi.org/10.1002/9781119011071.iemp0285

Bruns, A. (2021). Echo chambers? Filter bubbles? The misleading metaphors that obscure the real problem. In *Hate speech and polarization in participatory society* (pp. 33-48). Routledge.

Buckingham, D. (2019). Teaching media in a 'post-truth' age: Fake news, media bias and the challenge for media/digital literacy education. *Culture and Education*, 31(2), 213–231.

Center for Countering Digital Hate. (2024). *STAR Framework: Reducing the algorithmic amplification of hate and misinformation online*. https://counterhate.com/star

Churches, A. (2009). Taxonomía de Bloom para la era digital.

Comisión Europea. (2024). *Reglamento (UE) 2024/1689 por el que se establecen normas armonizadas sobre inteligencia artificial*. https://eur-lex.europa.eu/legal-content/ES/TXT/?uri=CELEX%3A32024R1689

Compton, J., & Pfau, M. (2009). Spreading inoculation: Inoculation, resistance to influence, and word-of-mouth communication. Communication Theory, 19(1), 9–28. https://doi.org/10.1111/j.1468-2885.2008.01330.x

Crespo Martínez, I., Melero-López, I., Mora Rodríguez, A., & Rojo Martínez, J. M. (2024). Política, uso de medios y polarización afectiva en España.

Damborrenea, R. G. (2000). Diccionario de falacias.

Damborrenea, R. G. (2011). Uso de razón: El arte de razonar, persuadir, refutar. Ediciones Uso de Razón.

Del-Fresno-García, M. (2019). Desórdenes informativos: sobreexpuestos e infrainformados en la era de la posverdad. *Profesional De La información*, *28*(3). https://doi.org/10.3145/epi.2019.may.02

Emsley, R. (2023). ChatGPT: These are not hallucinations – they're fabrications and falsifications. Schizophrenia, 9, Article 52. https://doi.org/10.1038/s41537-023-00379-4

Feliz-Murias, T. y Leví Orta, G. (2013). El humor como estrategia de motivación.

Fernández, P. C. (2023). Efectos de la sobrecarga de información en el comportamiento del consumidor de noticias: El doomscrooling. VISUAL REVIEW. International Visual Culture Review/Revista Internacional De Cultura Visual, 14(1), 1-11.

Freire, P. (1975). Pedagogía del oprimido (14.ª ed.). Siglo XXI Editores. https://isbn.org/9788432301841

Hao, C., Uusitalo, S., Figueroa, C., Smit, Q. T., Strange, M., Chang, W. T., ... & de Boer, M. H. (2025). A human-centered perspective on research challenges for hybrid human artificial intelligence in lifestyle and behavior change support. *Frontiers in Digital Health*, *7*, 1544185.

Garg, R., Han, J., Cheng, Y., Fang, Z., & Swiecki, Z. (2024). Automated discourse analysis via generative artificial intelligence. Proceedings of the 14th Learning Analytics and Knowledge Conference (LAK '24), 814–820. https://doi.org/10.1145/3636555.3636879

Gil-Quintana, J., Osuna-Acedo, S., Limaymanta, C. H., & Romero-Riaño, E. (2023). Análisis bibliométrico de artículos sobre innovación educativa en educación a distancia: Un reto para la pedagogía crítica y la educación mediática. *American Journal of Distance Education*, 37(4), 308–326. https://doi.org/10.1080/08923647.2023.2241715

Giovanola, B., & Granata, P. (2024). Ethics for human-centered education in the age of AI. En F. Spigarelli, L. Kempton, & L. Compagnucci (Eds.), *Entrepreneurship and digital humanities* (pp. 96–109). Edward Elgar Publishing.

Giroux, H. A. (1995). *Teoría y resistencia en educación*. Siglo XXI.

Granata, P. (2024). A chatbot for a thought: The flower of evil has bloomed (60 years later). *H-ermes. Journal of Communication, 26*, 23–36. https://doi.org/10.1285/i22840753n26p23

Islas, O., Cortés, F. G., & Urrutia, A. A. (2024). Una mirada a los riesgos y amenazas de la inteligencia artificial, desde la Ecología de los Medios. Comunicar: Revista Científica de Comunicación y Educación, (79), 1-9.

Jandrić, P. (2023). Postdigital. In: Jandrić, P. (eds) Encyclopedia of Postdigital Science and Education . Springer, Cham. https://doi.org/10.1007/978-3-031-35469-4_23-1

Jeon, Y., Kim, B., Xiong, A., Lee, D., & Han, K. (2021). Chamberbreaker: Mitigating the echo chamber effect and supporting information hygiene through a gamified inoculation system. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 1-26.

Kadushin, C. (2011). Understanding social networks: Theories, concepts, and findings. Oxford University Press.

Kahneman, D. (2011). *Pensar rápido, pensar despacio* (J. Chamorro Mielke, Trad.). Debate.

Lelkes, Y., Sood, G., & Iyengar, S. (2017). The hostile audience: The effect of access to broadband Internet on partisan affect. *American Journal of Political Science*, 61(1), 5–20. https://doi.org/10.1111/ajps.12237

Leyens, J. P., Demoulin, S., Vaes, J., Gaunt, R., & Paladino, M. P. (2007). Infra-humanization: The wall of group differences. *Social Issues and Policy Review*, *1*(1), 139-172.

Lévy, P. (2004). *Inteligencia colectiva: Por una antropología del ciberespacio* (F. Martínez Álvarez, Trad.). Organización Panamericana de la Salud. (Trabajo original publicado en 1994 como *L'intelligence collective: Pour une anthropologie du cyberespace*).

McGuire, W. J. (1964). Inducing resistance to persuasion: Some contemporary approaches. Advances in Experimental Social Psychology, 1, 191-229.

McIntyre, L. (2018). Post-truth. MIT Press.

Molenaar, I. (2022). *Towards hybrid human-AI learning technologies*. European Journal of Education, 57(4), 556–571. https://doi.org/10.1111/ejed.12525

Mollick, E. R., & Mollick, L. (2022). New modes of learning enabled by AI chatbots: Three methods and assignments. *Available at SSRN 4300783*.

Mollick, E. R., & Mollick, L. (2023). Using AI to implement effective teaching strategies in classrooms: Five strategies, including prompts. *The Wharton School Research Paper*.

Mollick, E., & Mollick, L. (2023). Assigning AI: Seven approaches for students, with prompts. *arXiv preprint arXiv:2306.10052*.

Mollick, E., & Mollick, L. (2024). Instructors as innovators: A future-focused approach to new AI learning opportunities, with prompts. *arXiv preprint arXiv:2407.05181*.

Observatorio Español del Racismo y la Xenofobia (OBERAXE). (2022). *El discurso de odio en redes sociales: análisis, detección y respuesta institucional*. Ministerio de Inclusión, Seguridad Social y Migraciones. https://www.inclusion.gob.es/oberaxe/es/publicaciones/documentos/el-discurso-del-odio-en-redes-sociales

Osuna-Acedo, S., Frau-Meigs, D., & Marta-Lazo, C. (2018). Educación mediática y formación del profesorado. Educomunicación más allá de la alfabetización digital. Revista interuniversitaria de formación del profesorado, 32(1), 29-42.

Pariser, E. (2011). The filter bubble: What the Internet is hiding from you. New York: Penguin Press.

Pariser, E. (2017). El filtro burbuja: Cómo la web decide lo que leemos y lo que pensamos (1.ª ed.). Taurus.

Pedroche-Santoveña, I. (2024). Interacciones y" gatekeepers" en la formación de cámaras de eco en X: caso de estudio# garzon. In *La comunicación ante el reto de las inteligencias artificiales, innovación, investigación y transferencias* (pp. 308-335). Dykinson.

Phoenix, J., & Taylor, M. (2024). Prompt engineering for generative AI: Future-proof inputs for reliable AI outputs. O'Reilly Media.

Racioppe, B. (2025). Postdigitalidad y memética. Reflexiones educativas sobre la generación de imágenes con IA a partir del trending topic "Cuando el genio malinterpreta nuestro deseo". Communiars. Revista de Imagen, Artes y Educación Crítica y Social, 13, 75-94. https://dx.doi.org/10.12795/Communiars.2025.i13.05

Rodríguez-Pérez, A., & Betancor, V. (2023). Infrahumanization: a restrospective on 20 years of empirical research. *Current Opinion in Behavioral Sciences*, *50*, 101258.

Roozafzai, Z. S. (2024). Unveiling power and ideologies in the age of algorithms: Exploring the intersection of critical discourse analysis and artificial intelligence. Qeios. https://doi.org/10.32388/60YE02

Salas, E. (2018). Influencia de los 11 principios de Joseph Goebbels en la campaña política de Donald Trump. Revista Caribeña de Ciencias Sociales.

Sardi, J., Candra, O., Yuliana, D. F., Yanto, D. T. P., & Eliza, F. (2025). How Generative AI Influences Students' Self-Regulated Learning and Critical Thinking Skills? A Systematic Review. *International Journal of Engineering Pedagogy*, *15*(1).

Scott, E. E., Rivale, S. D., & Nelson, M. A. (2020). Design-based research: A methodology to extend and enrich biology education research. CBE—Life Sciences Education, 19(2), es11. https://doi.org/10.1187/cbe.19-11-0252

Sunstein, C. R. (2001). *Republic.com*. Princeton University Press.

Sunstein, C. R. (2017). #Republic: Divided democracy in the age of social media. Princeton University Press.

Sperber, D., & Wilson, D. (2004). La teoría de la relevancia. Revista de Investigación Lingüística, 7(1), 237-288.

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. En W. G. Austin & S. Worchel (Eds.), The social psychology of intergroup relations (pp. 33-47). Brooks/Cole.

Tondello, G. F., Wehbe, R. R., Diamond, L., Busch, M., Marczewski, A., & Nacke, L. E. (2016). The Gamification User Types Hexad Scale. En CHI PLAY '16: Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play (pp. 229–243). Association for Computing Machinery. https://doi.org/10.1145/2967934.2968082

Törnberg, P. (2018). Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLoS one*, *13*(9), e0203958

Törnberg, P. (2022). How digital media drive affective polarization through partisan sorting. *Proceedings of the National Academy of Sciences*, 119(42), e2207159119. https://doi.org/10.1073/pnas.2207159119

Törnberg, P., Andersson, C., Lindgren, K., & Banisch, S. (2021). Modeling the emergence of affective polarization in the social media society. *PLOS ONE*, 16(10), e0258259. https://doi.org/10.1371/journal.pone.0258259

Törnberg, P., & Törnberg, A. (2024). Inside a White Power echo chamber: Why fringe digital spaces are polarizing politics. New Media & Society, 26(8), 4511-4533.

UNESCO. (2021). *El discurso del odio en redes sociales: Manual para educadores*. https://unesdoc.unesco.org/ark:/48223/pf0000379829

Van Dijck, J. (2016). *La cultura de la conectividad: Una historia crítica de las redes sociales* (H. Salas, Trad.) [Versión Kindle]. Siglo XXI Editores.

van Dijk, T. A. (1993). Principles of critical discourse analysis. *Discourse & Society*, 4(2), 249–283. https://doi.org/10.1177/0957926593004002006

van Dijk, T. A. (2015). Critical discourse analysis. En D. Tannen, H. E. Hamilton, & D. Schiffrin (Eds.), The Handbook of Discourse Analysis (pp. 466-485). Wiley Blackwell.

van Dijck, J. (2020). Seeing the forest for the trees: Visualizing platformization and its governance. *New Media & Society*, 23(9), 2801–2819. https://doi.org/10.1177/1461444820940293 *(Original work published 2021)*

Vargas González, C. A., & Quintero Carvajal, D. P. (2023). Aportes de la mayéutica socrática a la educación dialógica. Sophia, colección de Filosofía de la Educación, 35, 73-96. https://doi.org/10.17163/soph.n35.2023.02

Weiss, A. P., Alwan, A., Garcia, E. P., y García, J. (2020). Surveying fake news: Assessing university faculty's fragmented definition of fake news and its impact on teaching critical thinking. International Journal For Educational Integrity, 16(1), 1-30. Doi: http://dx.doi.org/10.1007/s40979-019-0049-X