Tsafiqui – Revista Científica en Ciencias Sociales
Vol. 16 Nº 1, 2026 | ISSN-E: **2602-8069** - ISSN-L: **1390-5341** | Universidad UTE
https://revistas.ute.edu.ec/index.php/tsafiqui/index



Inteligencia artificial en condiciones de ambigüedad radical: modelos mentales y emociones públicas

Artificial Intelligence Under Deep Uncertainty: Mental Models and Public Emotion

https://doi.org/10.29019/tsafiqui.v16i1.1694

Dante Avaro. Consejo Nacional de Investigaciones Científicas y Técnicas, Argentina

RESUMEN

Aquí se propone un marco analítico para comprender cómo se forman las expectativas públicas sobre la inteligencia artificial (IA) bajo condiciones de ambigüedad conceptual radical. A partir de datos secundarios de la Chapman Survey of American Fears, Pew Research Center (2021-2025) y YouGov (2025), el trabajo examina juicios y emociones públicas en Estados Unidos respecto a la posibilidad de que la IA sea o llegue a ser "más inteligente que las personas". Para ello, se proponen y articulan tres mecanismos de procesamiento: (MP1) heurísticas afectivas moduladas por la familiaridad, que distinguen entre miedos abstractos/existenciales y preocupaciones concretas; (MP2) anclaje en el consenso experto percibido, filtrado por sesgos mediáticos e identitarios; y (MP3) construcción de modelos mentales, mediante los cuales legos y expertos definen de manera divergente qué cuenta como "IA", "inteligencia" y "superioridad". El análisis exploratorio identifica cinco patrones: heterogeneidad estructurada de expectativas, transformación del miedo con la experiencia, brecha sistemática entre consenso experto real y percibido, divergencias en la categorización de la automatización laboral y disminución diferenciada de la confianza en la IA a medida que aumenta la experiencia de uso. Estos hallazgos sugieren que las actitudes hacia la IA no pueden explicarse solo por la exposición o la información disponible, sino por la interacción entre afectos, señales de autoridad epistémica y modelos mentales que operan sobre un objeto intrínsecamente ambiguo.

ABSTRACT

This article proposes an analytical framework for understanding how public expectations about artificial intelligence (AI) emerge under deep uncertainty. Drawing on secondary data from the Chapman Survey of American Fears, Pew Research Center (2021-2025), and YouGov (2025), the study examines public judgments and emotions in the United States regarding whether AI is—or may become— "more intelligent than people". It specifies three information-processing mechanisms: (MP1) affective heuristics modulated by familiarity, which distinguish abstract/existential fears from concrete concerns; (MP2) anchoring in perceived expert consensus, filtered through media and identity-based biases; and (MP3) the construction of mental models through which laypeople and experts draw divergent boundaries around what counts as "AI," "intelligence," and "superiority." The exploratory analysis identifies five patterns: structured heterogeneity of expectations, experience-dependent shifts in fear, a systematic gap between actual and perceived expert consensus, divergences in the categorization of workplace automation, and differential declines in trust in AI as usage experience increases. Taken together, these findings suggest that attitudes toward AI cannot be explained solely by exposure or available information, but by the interplay of affect, signals of epistemic authority, and mental models operating on an intrinsically ambiguous object.

PALABRAS CLAVE | KEYWORDS

inteligencia artificial, miedo, afectividad, comportamiento, modelos mentales, EE. UU.
artificial intelligence, fear, emotions, behavior, mental models, USA.

Recibido: 14/11/2025 - **Aceptado**: 11/12/2025 - **Publicado**: 01/01/2026 | Páginas: 27-47

CC-BY 4.0

INTRODUCCIÓN

Durante estos últimos años se evidencia, al interior de la gran familia de las ciencias sociales, una revitalización de estudios sobre geopolítica. Dentro de estos destacan aquellos que abordan cuestiones que caen bajo el amplio paraguas denominado "geopolítica del miedo". Esta etiqueta ya no refiere solo a los clásicos conflictos interestatales, sino que explora la difusión ubicua del miedo a escala planetaria, algo así como una territorialización cultural del miedo que atraviesa a múltiples y diferentes espacios societales. Emerge, mediante este último recorte, un asunto novedoso: la inteligencia artificial (IA) y su relación con aquellos miedos difusos ya mencionados.

Las preocupaciones acerca del uso de la IA en la geopolítica del miedo son múltiples, aunque algunos pueden ser clasificados como convencionales y otros como novedosos. Entre los primeros, cabe destacar el uso de la IA en eventos bélicos presentes e hipotéticos, en los conflictos comerciales —especialmente entre EE. UU. y China— y sus efectos macroeconómicos a escala mundial y, para cerrar esta provisoria lista, los efectos de la despiadada lucha por liderar la industria de la IA, que incluye, por supuesto, el control sobre los estándares internacionales de producción en dicha industria. Entre los segundos, las inquietudes e intranquilidades no se reducen a observar cómo la IA está modificando las estructuras de gobernanza (pública y privada), sino también si de esta frenética difusión y adopción ubicua de la IA surgirá un nuevo esquema de legitimidad para las estructuras de poder.

Por tanto, el asunto a comprender ya no es únicamente qué hacen o dejan de hacer los Estados y las corporaciones con la IA, sino que incluye el problema de cómo la IA es imaginada, temida y evaluada por la población. Si la premisa que sugiere que en la geopolítica del miedo las sociedades están atravesadas por espacios culturales de afectos es correcta, entonces importa estudiar no solo las estrategias implementadas desde arriba (p. ej., Estado), sino también las condiciones cognitivas y emocionales disponibles en la población, bajo las cuales aquellas estrategias pueden operar (*v. gr*. qué objetos de miedo se estabilizan, qué temores se vuelven verosímiles, qué beneficios se aceptan como plausibles y qué ambivalencias persisten). Además, en el caso de la IA, las orientaciones que se figura la población no se forman frente a un objeto plenamente definido.

Por el contrario, la IA representa un objeto cambiante, difuso, incierto y envuelto en una disputa semántica que tiende a amplificar la dependencia de la población de marcos interpretativos, señales de autoridad epistémica y narrativas mediáticas, la mayoría de las veces en tensión. Por eso, intentar, como se hará en este trabajo, reconstruir los juicios y emociones públicas ante la IA ofrece una provisoria ventana sociopolítica exploratoria para observar uno de los temas centrales del acontecer contemporáneo.

Este trabajo, conforme a lo expuesto, inicia con una pregunta general: ¿qué juicios y emociones despierta, entre los humanos, la existencia de la IA? Para explorarla, indago cómo expertos y públicos no especializados están reorganizando sus intuiciones sobre la IA, es decir, qué temores despierta, qué peligros concretos le atribuyen, qué beneficios consideran plausibles y cómo esas respuestas se combinan en una mezcla estable de miedo, inquietud, familiaridad y ambivalencia al pensar la convivencia futura entre sistemas de IA y vida humana. La sección 2 presenta resultados de encuestas recientes que permiten identificar patrones de miedo, entusiasmo y ambivalencia.[1] La sección 3 propone un esquema analítico mínimo —afectos, modelos mentales, juicios evaluativos— para ordenar esas respuestas sin sobredimensionar su alcance. En la sección 4 utilizo ese esquema para interpretar los datos secundarios de YouGov y del Pew Research Center. El cierre delimita qué puede sostenerse con esta evidencia, qué permanece abierto y qué restricciones metodológicas exigen prudencia.

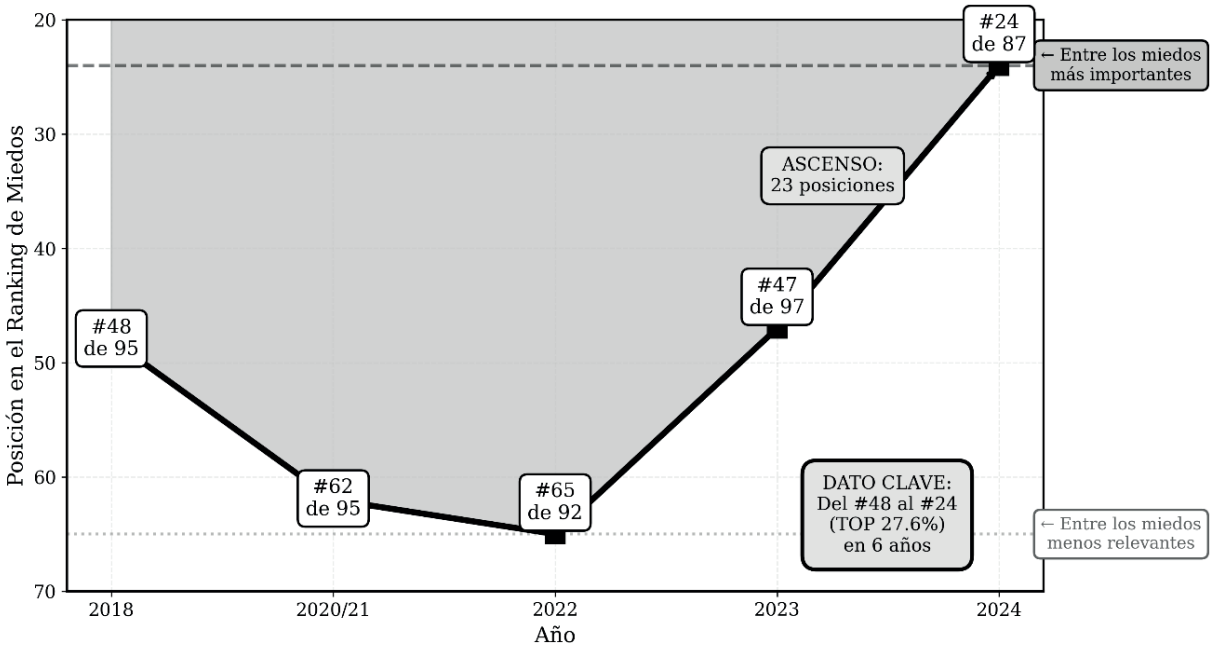
1. Esta investigación requirió hallar datos de acceso libre de a) al menos dos olas sobre preguntas relacionadas con miedo e IA, b) diferencias de percepción en torno a la IA entre expertos y público en general y c) al menos dos olas sobre miedos personales. La búsqueda arrojó que YouGov (2024 y 2025), Pew Research Center (2025) y la Encuesta Chapman sobre Miedos Estadounidenses satisfacían, respectivamente, la concurrencia de los tres criterios; y solo disponible para el caso estadounidense. Hay que aclarar que los datos de Pew Research 2023 no son de una única encuesta, sino una compilación de varios estudios que incluyen: a) una encuesta de agosto de 2023 sobre conciencia general de IA, b) una encuesta de diciembre de 2022 sobre usos específicos de IA, c) una encuesta de marzo de 2023 sobre ChatGPT, d) una encuesta de otoño de 2023 sobre adolescentes y ChatGPT y e) varias otras encuestas sobre temas específicos (salud, trabajo, vehículos autónomos). Por otra parte, si bien es cierto que la International Social Survey Programme, Gallup World Polls y Eurobarometer confeccionan encuestas sobre riesgos, no se enfocan en miedos personales. Finalmente, aunque YouGov es considerada uno de los proveedores más rigurosos dentro del espacio de encuestas en línea no probabilísticas, los expertos enfatizan que ninguna metodología de encuesta constituye un estándar de oro único y que se debe ejercer cautela interpretativa, especialmente con hallazgos sobre subgrupos pequeños o temas sensibles. Y el caso que aquí nos ocupa no es la excepción.

Por último, cerrando esta introducción, conviene señalar que esta propuesta se apoya en tres fuentes; cada una captura una dimensión distinta del problema y, en conjunto, permiten —sin acceder a microdatos brutos— triangular patrones. La Encuesta Chapman aporta una serie de olas adecuada para observar cambios temporales en afectos asociados a la IA (especialmente en el ámbito laboral). Pew Research Center permite comparar actitudes del público con percepciones/atribuciones sobre el consenso experto. YouGov, por su parte, ofrece ítems que obligan a los encuestados a pronunciarse sobre un evento altamente ambiguo (*v. gr.* "IA más inteligente que las personas"), lo cual es central para el argumento sobre ambigüedad radical. Por último, los indicadores y tablas se seleccionan con base en tres criterios. Primero, correspondencia teórica mínima con afectos, señales de autoridad epistémica y construcción del objeto de juicio; segundo, comparabilidad (mínima) en reportes públicos; y tercero, relevancia semántica para el asunto de la ambigüedad radical, es decir, preguntas cuya formulación exige "construir" qué se está evaluando, no solo estimar probabilidades.

EMOCIONES HACIA LA INTELIGENCIA ARTIFICIAL: DATOS Y CONTEXTO

El análisis del miedo a la IA en el ámbito laboral, llevado a cabo por la Encuesta Chapman sobre Miedos Estadounidenses, revela un cambio llamativo en la percepción pública estadounidense entre 2018 y 2024. Durante los primeros años de la serie (2018-2022), este miedo se mantuvo relativamente estable, fluctuando entre el 27.5 % y un 30.7 % de la población, ocupando posiciones intermedias en la lista general de miedos (*vid.* gráfico 1). Sin embargo, a partir de 2023 se observa un cambio de valoración pública que coincide temporalmente con el lanzamiento de ChatGPT en noviembre de 2022. Así, entre 2022 y 2024, el porcentaje de estadounidenses que temen que la IA reemplace a las personas en sus empleos prácticamente se duplicó, saltando de un 27.6 % a un 46.8 %.

Gráfico 1
El ascenso del miedo a la IA (EE. UU. 2018-2024)

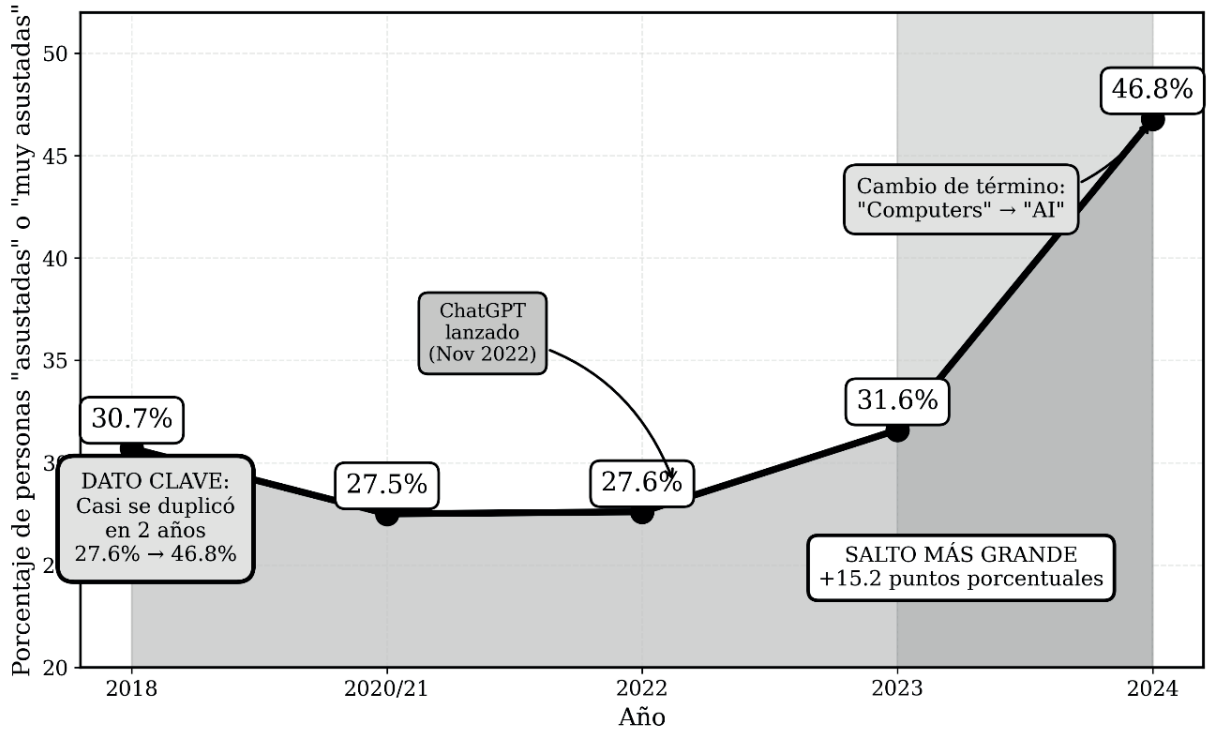


Nota. Encuesta Chapman sobre Miedos estadounidenses.

Durante 2024, la encuesta Chapman sustituyó el fraseo de "*computers replacing people in the workforce*" (ola 9 de 2023) por "*AI replacing people in the workforce*" (ola 10 de 2024). Este cambio puede ser interpretado no

solo como una evolución semántica, sino también como el reconocimiento de que la IA se ha convertido en una categoría conceptual distintiva y amenazante en el imaginario social. En 2024, un 46.8 % de los encuestados declaró tener miedo o mucho miedo a que la IA sustituya a las personas en el trabajo, una proporción cercana a la registrada para el miedo a la contaminación del aire (47.0 %) y ligeramente superior al miedo a tiroteos masivos (45.9 %). Este porcentaje supone un aumento de unos 15 puntos respecto de 2023, posicionándose como uno de los miedos que más han aumentado en los últimos años de la encuesta.

Gráfico 2
Miedo a ser desplazado laboralmente por la IA (EE. UU. 2018-2024)



Nota. Encuesta Chapman sobre Miedos estadounidenses

El Pew Research Center, por otra parte, llevó a cabo dos estudios relacionados con la IA, que abarcan varios temas, pero indagan específicamente en las expectativas públicas y emociones que existen entre la relación humano e IA. El primer estudio, publicado en 2023, resume varias encuestas llevadas a cabo entre 2021 y 2023 (Faverio y Tyson, 2023). El segundo, publicado en 2025, es el resultado de una encuesta única que compara expectativas sobre la IA entre el público y los expertos (McClain *et al.*, 2025). Estos estudios muestran un patrón que admite la siguiente interpretación: existe un miedo generalizado y creciente hacia la IA, que se manifiesta principalmente en cinco áreas. Estas son: a) pérdida de empleos y seguridad económica, b) pérdida de conexión humana y relaciones sociales, c) desinformación y mal uso de datos personales, d) velocidad de implementación de la IA sin suficientes salvaguardas (especialmente en el campo de la salud) y e) falta de regulación gubernamental efectiva. La anterior interpretación permite exponer una paradoja: quienes menos conocen la IA (el público) le temen más, mientras que los expertos mantienen mayor optimismo de cara al futuro.

En la misma línea de indagación, la empresa YouGov llevó a cabo, entre el público estadounidense y durante 2024 y 2025, dos encuestas sobre IA. En la última se replicó la siguiente pregunta: "¿Qué probabilidad cree que hay de que la inteligencia artificial (IA) llegue a ser más inteligente que las personas?"[2] La distribución de respuestas muestra una llamativa heterogeneidad (*vid.* tabla 1).

2. Encuesta llevada a cabo con público estadounidense durante el 5-7 de marzo de 2025. Pregunta: "How likely do you think it is that artificial intelligence (AI) will eventually become more intelligent than people?" (Sección 7 titulada: "AI Intelligence"), YouGov (2025).

Tabla 1. Distribución de expectativas sobre capacidades de la IA		
Categoría de Respuesta	**Y es:**	**Porcentaje**
Ya es más inteligente	Y_1	13 %
Muy probable que ocurra	Y_2	25 %
Algo probable que ocurra	Y_3	22 %
No muy probable	Y_4	15 %
Nada probable	Y_5	9 %
No está seguro	Y_6	17 %

Nota. YouGov (2025). Los porcentajes no suman 100 por redondeo (*vid.* YouGov, 2025). Las respuestas Y_1 e Y_6 no son semánticamente unívocas; sus posibles lecturas se discuten a lo largo del texto.

A partir de estos datos, surgen tres características analíticas.

Primera. El 13 % de los encuestados (*vid.* tabla 1) que eligen que "ya es más inteligente" es un dato que sugiere: (a) que los respondentes tienen una creencia genuina de que la inteligencia artificial general ha sido alcanzada; sin embargo, también puede ser (b) que tengan una confusión entre capacidades estrechas o específicas de la IA (sistemas que superan a humanos en tareas específicas) con la inteligencia general,[3] o (c) que hayan usado marcos interpretativos diferentes para el término "inteligente". Dado que no hay un consenso experto sobre la existencia de la AGI —aunque sí muchas advertencias serias sobre su inminente llegada, *cfr.* Yudkowsky y Soares (2025)—, este patrón de respuesta demanda mayor reflexión.

Segunda. El 17 % que expresa "no está seguro" representa un valor sustancial —uno de cada seis encuestados declina ofrecer un juicio— que requiere una interpretación cuidadosa (*vid*. tabla 1). Si bien la literatura sobre respuestas de "no opinión" muestra que no existe una cifra "típica" estándar, en un experimento clásico, sin embargo, Krosnick *et al*. (2002) encuentran que, cuando se ofrece explícitamente la opción de "no opinión", las tasas de "no sé" oscilan en torno al 6-13 %, mientras que, al omitirla, caen a valores cercanos al 1-2 % en los mismos ítems. En estudios sobre temas poco familiares o incluso ficticios, hasta ≈30 % de los entrevistados llegan a expresar una opinión si no se facilita la opción "no sé", lo que refleja tanto inferencia razonable como actitudes poco consolidadas. En suma, el porcentaje de "no sé" varía sustantivamente según el contenido y, sobre todo, según el formato de respuesta, de modo que conviene evitar aplicar umbrales fijos sin contexto (Krosnick *et al.*, 2002).

A la luz de este marco, el 17 % que responde "no está seguro" constituye un caso analíticamente distinto. A falta de medidas directas de modelos mentales, una interpretación plausible es que este segmento refleja sobre todo la ausencia de esquemas operativos para pensar la IA, más que un reconocimiento experto de ambigüedad. Así, se puede suponer que no se trata tanto de individuos que suspenden deliberadamente el juicio porque perciben límites epistémicos, sino de encuestados que aún no disponen de un mapa mínimo del problema, es decir, no saben bien qué cuenta como "IA", qué significa que sea "más inteligente que las personas" ni qué implicaciones prácticas tendría. Otras interpretaciones —como la exposición a argumentos contradictorios o la percepción de que la pregunta está subespecificada— no pueden descartarse por completo, pero el comportamiento posterior de la serie sugiere que el componente dominante es la falta de (un) modelo mental. Aquí propongo que esta lectura es, aunque no pueden descartarse otras, coherente con la reducción del "no sé" a medida que aumentan la experiencia y la exposición a la IA (*vid. infra* §.16.).

Tercera. Las encuestas de YouGov muestran que, mientras la familiaridad con la IA ha aumentado (reflejada por la difusión y adopción ubicua de ChatGPT y herramientas similares), el miedo al riesgo existencial (*i.e.,*

3. Los acrónimos utilizados en idioma inglés para *artificial general intelligence*, *artificial narrow intelligence* y *artificial superintelligence* son AGI, ANI y ASI, respectivamente. A lo largo del texto se utilizan estos acrónimos para referenciar la IA general, la IA específica y la superinteligencia artificial.

miedo a que la IA cause la extinción humana) ha disminuido[4]; sin embargo, la preocupación sobre impactos concretos es relevante (p. ej., en cuanto a la preocupación porque la IA genere un desplazamiento laboral, *vid.* Tabla 2). Estos datos, analizados en conjunto, admiten la interpretación de un patrón que resulta inconsistente con modelos simples de heurística afectiva que predicen relaciones monotónicas entre exposición y ansiedad (*cfr.* Slovic *et al.*, 2004). Bajo este marco interpretativo, lo anterior se puede exponer así: si la familiaridad reduce el miedo, entonces, ¿por qué crece la preocupación concreta?

La interpretación de los datos subyacentes en las primeras dos caracterizaciones conforma un patrón que apunta hacia un aspecto crucial: la ambigüedad radical o de segundo orden. Hay ambigüedad radical cuando el objeto del juicio está subespecificado en dimensiones clave (p. ej., dominio, significado del predicado u horizonte temporal), de modo que personas distintas pueden estar respondiendo a objetos de juicio diferentes. Lo que significa que la ambigüedad no solo opera sobre las probabilidades o resultados (aspecto central en la literatura sobre el juicio bajo incertidumbre), sino sobre el constructo epistémico en sí mismo.[5] En concreto, si bien la pregunta formulada por YouGov interroga sobre la probabilidad de que la IA "llegue a ser más inteligente que las personas", hay que destacar que resulta ambigua en tres aspectos: a) en el alcance del dominio, b) en la delimitación del significado de "inteligente" y, finalmente, c) en la especificación temporal. Esta ambigüedad no es, *per se,* un defecto metodológico en la encuesta de YouGov, sino que refleja cómo existe la pregunta en el discurso y la opinión pública. Así, la encuesta captura expectativas tal como se forman en su contexto "natural", es decir, bajo condiciones de ambigüedad radical. Sin embargo, este asunto importa porque la mayoría de los modelos de juicio bajo incertidumbre asumen que el objeto del juicio es claro (Tversky y Kahneman, 1974), variando solo la información sobre las probabilidades o los resultados (para un enfoque crítico, *cfr.* Gigerenzer y Goldstein, 1996; Slovic *et al.*, 2004). Pero la pregunta "¿llegará la IA a ser más inteligente que las personas?" presenta un problema cualitativamente diferente: el evento mismo es conceptualmente ambiguo. Los encuestados primero, antes de poder asignar una probabilidad, deben construir una interpretación de qué se les pregunta.[6]

Una forma de abordar el anterior problema es a través de los modelos de procesamiento dual, ampliamente conocidos por los aportes de Kahneman (2011) y por las especificaciones de "autonomía" vs. "*decoupling*" más "memoria de trabajo" propuestas por Evans y Stanovich (2013). Estos modelos proponen que la cognición humana opera a través de dos sistemas, simplificando, un modelo de procesamiento intuitivo y otro reflexivo (*cfr.* Sloman, 1996). La fortaleza del primero radica en que les permite a las personas tomar decisiones rápidas con una carga cognitiva mínima, mientras que la fortaleza del segundo es que, al poder realizar un razonamiento abstracto y deliberar, puede tomar control sobre el primero. Por otra parte, la mayor debilidad del primero es que es un sistema de procesamiento susceptible a sesgos sistemáticos, mientras que la debilidad del segundo es que es cognitivamente costoso, limitado en capacidad y frecuentemente perezoso (se recuesta sobre la disponibilidad inmediata que ofrece el primero). Este enfoque ha sido aplicado productivamente a la percepción de riesgo tecnológico (*cfr.* Slovic *et al.*, 2004; Siegrist *et al.*, 2005). Sin embargo, aplicarlo directamente a las expectativas sobre IA enfrenta tres limitaciones.

4. Los datos de las encuestas de YouGov muestran que los respondentes "muy/algo preocupado" pasaron del 39 % en 2024 al 37 % en 2025.

5. La ambigüedad en la etiqueta "ambigüedad radical" no apunta al clásico asunto de restringir o liberar a los conceptos de un significado único, sino que señala un contexto en donde la incertidumbre no recae (solo) sobre las probabilidades o resultados, sino sobre el constructo epistémico que se está evaluando. En consecuencia, el evento es conceptualmente ambiguo y los individuos deben construir una interpretación o aplicar algún modelo mental antes de poder asignar probabilidades. Así, ambigüedad radical, en conjunto, se refiere al contexto de incertidumbre que rodea a la IA no solo en el plano óntico (*i.e.*, desacuerdos sobre características o rasgos sobre él, para usar la frase de Bunge, "moblaje del mundo"), sino —de forma más decisiva— al plano ontológico, es decir, sobre qué tipo de entidad se supone que es la IA y qué cuenta como "inteligencia". Así, el problema que la etiqueta ambigüedad radical quiere enfatizar no es que los legos no puedan reconocer ejemplos de IA, sino que pueden carecer de los recursos cognitivos necesarios para delimitar el concepto, moverse a lo largo de sus dimensiones y comparar interpretaciones rivales. De allí que requieran construir modelos mentales parciales para enfrentar la pregunta.

6. Aquí los autoinformes sobre expectativas tecnológicas no se interpretan como pronósticos o juicios asertivos calibrados, sino como juicios subjetivos de plausibilidad y evaluación afectiva frente a un objeto subespecificado. Su validez, por tanto, es sociocognitiva, es decir, interesa qué interpretación construyen los individuos y qué señales usan para responder. Esto implica que el valor empírico de las respuestas no depende de "acertar" (o no) acerca de estadios futuros, sino de explorar cómo se forman expectativas públicas cuando el evento evaluado es conceptualmente ambiguo y disputado.

Estudios de percepción de riesgo tecnológico (Sjöberg, 2000; Fischhoff *et al.*, 1978) sugieren que la familiaridad con "nuevas" tecnologías tiende, en líneas generales, no solo a reducir el miedo inicial, sino que además produce una transformación o reconfiguración del miedo (un cambio en su carácter/tipología). Así, si bien la exposición inicial genera un miedo difuso y existencial (p. ej., juicios del tipo "la tecnología destruirá la sociedad"), la exposición continua, por otra parte, reduce ese temor abstracto, pero aumenta la conciencia de miedos específicos y concretos. Esto sugiere que es preciso distinguir tipos de afecto, no solo intensidad (una distinción que está ausente en el Sistema 1 estándar). Por tanto, se podría asumir provisoriamente que los afectos no producen resultados unitarios.

Los juicios sobre tecnologías emergentes están fundamentalmente mediados por el consenso experto percibido (*cfr.* Kahan *et al.*, 2011; Lewandowsky *et al.*, 2013). Cuando los individuos carecen de experticia técnica, infieren respuestas "correctas" de lo que perciben como consenso científico. Este proceso no es ni puro Sistema 1 (*i.e.,* no es el resultado de un proceso automático/afectivo) ni puro Sistema 2 (no resulta de un razonamiento analítico individual), sino un proceso social-epistémico distinto, que puede etiquetarse como deferencia a la autoridad epistémica.[7]

Un asunto más relevante aún que los anteriores es que la teoría del procesamiento dual no aborda el asunto de cómo los individuos categorizan y definen los constructos (epistémicos) ambiguos. La investigación sobre diferencias entre experto-novato (*vid.* Chi *et al.*, 1981) muestra que el proceso de categorización no es meramente un *input* que se introduce mecánicamente en los juicios, sino que es un proceso constructivo que varía sistemáticamente con el nivel de conocimiento que se tenga sobre el dominio en cuestión. Así, extendiendo el análisis al caso de la IA, podría ser que un experto que escucha la proposición "más inteligente que las personas" probablemente piensa "AGI con razonamiento general de nivel humano" (un constructo técnico con significado específico). Mientras que un novato, ante la misma frase, puede pensar en algo así como "sistemas que me impresionan con sus resultados" (colapsando, de este modo, ANI con AGI). Lo relevante es que estos son objetos de juicio distintos, no evaluaciones de probabilidad diferentes del mismo objeto. Este proceso —que etiqueto como construcción de modelos mentales— opera con una lógica alternativa tanto de las heurísticas afectivas (Sistema 1) como del razonamiento analítico (Sistema 2). En síntesis, la construcción de modelos mentales es el proceso mediante el cual los individuos resuelven la ambigüedad construyendo una interpretación, frecuentemente sin conciencia (total) de que existen alternativas (simplemente porque no pueden hacerlo).

Para comprender la formación de expectativas bajo ambigüedad conceptual radical, utilizo un esquema mínimo (no exhaustivo) de tres componentes o marcos de procesamiento, porque la evidencia disponible y el objeto de estudio obligan a distinguir tres asuntos diferentes. En primer lugar, cómo se valora el riesgo cuando falta comprensión técnica; segundo, bajo qué señales de autoridad epistémica se ancla la creencia; y tercero, cómo se construye el significado del evento bajo juicio.[8] A continuación, se presentan.

- Heurística afectiva modulada por familiaridad (en adelante MP1). El mecanismo central lo constituyen las reacciones emocionales que sirven (o pueden ser utilizadas) como *proxies* para juicios de probabilidad (*cfr.* Slovic *et al.*, 2004). Esto permite distinguir entre el miedo abstracto (pre-familiar, es decir, escenarios apocalípticos *ex ante*) y el miedo concreto (post-familiaridad, *i.e.,* amenazas específicas como *deepfakes*, pérdida de empleo, pérdida de la privacidad, etc.).
- Anclaje en el consenso experto percibido (en adelante MP2). Cuando los individuos enfrentan información técnica o científica que excede su comprensión (incertidumbre técnica), recurren racionalmente a señales sociales —ya sean las acciones de otros (cascadas) o las evaluaciones de expertos (consenso científico)— para inferir qué creencia o acción es correcta (Bromme y Goldman, 2014; Bikhchandani *et al.*, 1992). Esta estrategia no es, *per se,* un sesgo o error, sino una respuesta adaptativa y racional a las limi-

7. Conviene resaltar que muchas veces el consenso percibido diverge del consenso real (*vid.* Oreskes, 2004; Lewandowsky *et al.*, 2013 y Koehler, 2016). Sin embargo, es conveniente advertir, que el discurso público amplifica las voces expertas alarmistas sobre las optimistas (para más detalles *vid. infra* nota al pie 10).

8. Los componentes o marcos del esquema no son mutuamente excluyentes, pueden operar en paralelo y, con frecuencia, de forma secuencial. La interacción mínima se puede exponer del siguiente modo: MP3 fija el objeto de juicio (p. ej., "qué cuenta como IA más inteligente"), MP2 aporta anclajes sociales (qué se cree que piensan los expertos, filtrado por medios e identidades) y MP1 aporta la valencia y el atajo evaluativo (es decir, miedos abstractos vs. preocupaciones concretas moduladas por familiaridad).

taciones cognitivas inevitables frente a la complejidad del conocimiento científico moderno. Por tanto, bajo incertidumbre técnica, los individuos infieren la respuesta "correcta" a través del consenso científico percibido (Bromme y Goldman, 2014). La literatura especializada afirma que el consenso percibido puede apartarse del consenso real (*i.e.*, del estado del conocimiento científico) debido a la amplificación mediática selectiva y el filtrado cultural (Kahan *et al.*, 2012).

- Construcción de modelos mentales bajo ambigüedad (en adelante MP3). Antes de asignar probabilidades, los individuos construyen interpretaciones de qué significa (o puede significar) una "IA más inteligente que las personas" (*cfr.* Johnson-Laird, 1983; Chi *et al.*, 1981).

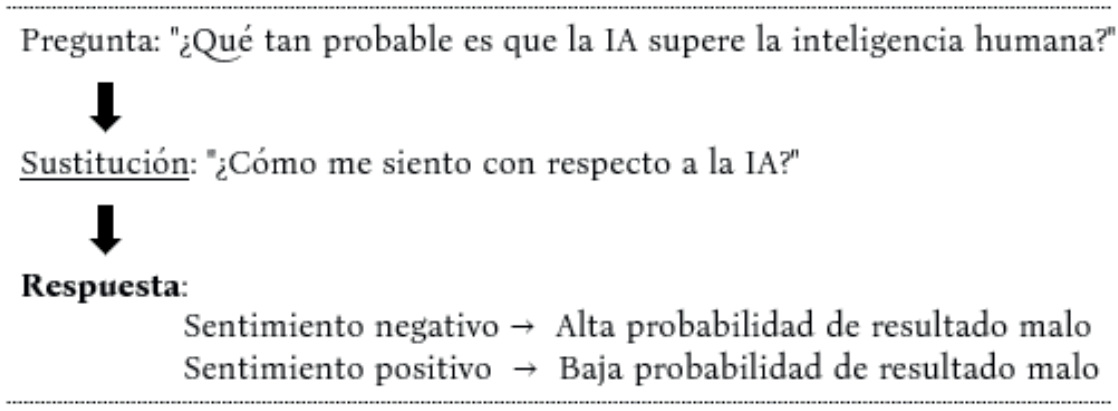
MICRO FUNDAMENTOS

Hace algunos años, Slovic y colaboradores (*vid.* Slovic *et al.*, 2004) aplicaron la teoría de procesamiento dual para explicar reacciones públicas a (nuevas) tecnologías. Propusieron una heurística afectiva que establece que cuando la información técnica es compleja o no disponible, las personas sustituyen la pregunta "¿qué tan riesgoso es X?" por una pregunta alternativa similar a "¿Cómo me siento sobre X?". El hallazgo es que los sentimientos y las emociones sirven como atajo informacional.[9] Aunque este aporte es muy valioso, aplicado directamente a las expectativas sobre IA, parece enfrentar tres limitaciones. Primera, el afecto no es monolítico. El "miedo a la IA" observado en 2024 es, a pesar de que la familiaridad aumentó, cualitativamente diferente del "miedo a la IA" observado en 2025. Lo que sugiere que es necesario discriminar y distinguir entre tipos de afecto. Segundo, el consenso experto importa, pero el público no accede a él directamente, sino a representaciones filtradas por medios y lentes culturales. Este proceso social-epistémico no encaja limpiamente en el Sistema 1 o el Sistema 2. Tercero, el objeto que se está juzgando es ambiguo. Antes de aplicar el Sistema 1 o el Sistema 2, los individuos deben construir una interpretación de qué significa la frase "IA más inteligente que las personas". Este proceso previo no es ni automático-afectivo ni controlado-analítico; es un proceso distinto de construcción conceptual.

Ahora, describo con más detalle los tres marcos (o componentes) de procesamiento cognitivo.

Fundamentación de MP1. Las personas, cuando son confrontadas con preguntas técnicas complejas, usan sus reacciones emocionales como un *proxy* informacional (*cfr.* Schwarz y Clore, 1983; Finucane *et al.*, 2000). El esquema 1 representa, de forma simplificada, dicho proceso.

Esquema 1



Nota. Slovic *et al.* (2004).

9. Esto se puede ilustrar mediante la divergencia entre expertos y público. Mientras los expertos usan evaluación analítica de riesgo (Sistema 2), el público usa afecto (Sistema 1).

Esto quiere decir que si la heurística afectiva fuera, para el problema aquí analizado, operativamente decisiva, entonces se podrían predecir dos escenarios hipotéticos.

> Ha: Los individuos con mayor miedo frente a la IA deberían asignar mayor probabilidad a que la "IA se vuelve más inteligente".
>
> Hb: Este efecto debería ser más fuerte para aquellos individuos con bajo conocimiento técnico (*i. e.,* en quienes dependen más del Sistema 1).

Sin embargo, estos escenarios son demasiado simples, ya que el análisis de cortes transversales repetidos de YouGov (2024-2025) revela un patrón empírico que requiere una mejor explicación.

Tabla 2. Cambios en el miedo relacionados con la IA

Tipo de Preocupación	2024	2025	Cambio (pp)	Interpretación
Abstracta/Existencial				
"IA causará extinción humana"	39 %	37 %	-2	Disminución
"Asustado" de IA (general)	~22 %	22 %	0	Estable
Concreta/Específica				
Expectativa por caída del empleo	57 %	N/A	—	
Preocupación por desplazamiento laboral	N/A	77 %	—	Nuevo, muy alto
Preocupación por *deepfakes*	N/A	85 %	—	Nuevo, muy alto
Erosión de privacidad	N/A	83 %	—	Nuevo, muy alto
Manipulación de comportamiento	N/A	80 %	—	Nuevo, muy alto
Confianza en la IA				
Información precisa	43 %	48 %	+5	Aumento
Decisiones éticas	21 %	23 %	+2	Aumento

Nota. YouGov (2024; 2025).

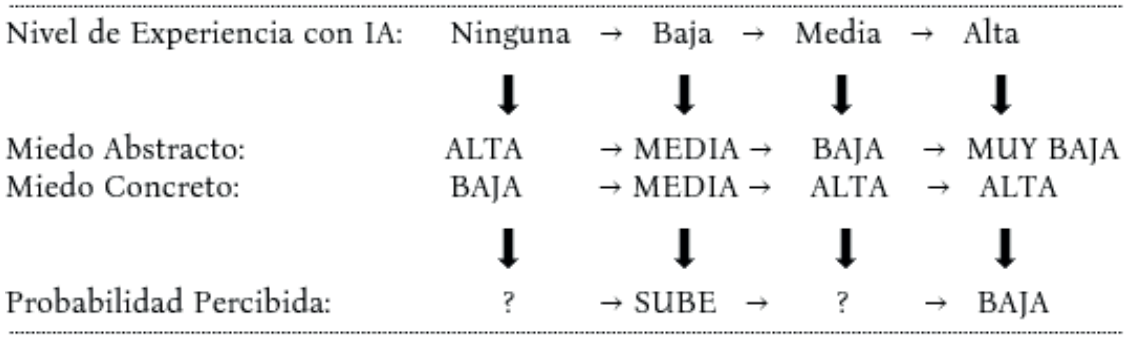
La comparación temporal en la tabla 2 revela un patrón diferenciado en la evolución de temores hacia la IA. Suponiendo que la familiaridad pública con la IA aumentó durante este período, las teorías estándar sobre heurística afectiva generan predicciones contrapuestas: la familiaridad debería reducir el miedo uniformemente (Fischhoff *et al.*, 1978), o bien, la mayor preponderancia debería aumentarlo uniformemente al incrementar la disponibilidad cognitiva del riesgo (Tversky y Kahneman, 1974). Los datos agregados, sin embargo, muestran un patrón más complejo. Por un lado, los temores abstractos o existenciales disminuyen o permanecen estables (temor a "extinción humana", sentirse "asustado" en general), mientras que en 2025 emergen preocupaciones concretas o específicas con niveles muy altos (desplazamiento laboral, *deepfakes*, erosión de privacidad, manipulación de comportamientos).

Este patrón es consistente con la hipótesis de que la familiaridad no reduce simplemente la intensidad del afecto hacia la IA, sino que transforma su objeto y modalidad, transitando de temores difusos y existenciales (Modo A) hacia preocupaciones calibradas sobre riesgos específicos y anclados en la experiencia (Modo B). Bajo esta interpretación, los datos sugieren dos modalidades diferenciadas de procesamiento afectivo: (a) el miedo abstracto o existencial, predominante en contextos de baja familiaridad, disminuye con la exposición repetida a aplicaciones concretas de IA; y (b) el miedo concreto o específico, emergente en contextos de mayor familiaridad, aumenta conforme los individuos identifican riesgos tangibles vinculados a usos concretos de la IA.

Este hallazgo, sin embargo, podría interpretarse, también, por (a) efectos de composición muestral (diferentes poblaciones en 2024 vs. 2025), (b) cambios en el contexto político-mediático que intensificaron preocupaciones específicas independientemente de la familiaridad, o (c) por diferencias metodológicas entre olas de encuesta. Sin embargo, la direccionalidad opuesta de los cambios (temores abstractos estables/descendentes vs. concretos/altos), su consistencia con marcos teóricos sobre procesamiento de riesgos tecnológicos (*cfr.* Slovic, 1987) y su coherencia con el efecto del conocimiento autoinformado hacen que la interpretación de transformación afectiva no pueda ser descartada rápidamente.

Lo anterior puede ser usado para establecer un punto de partida entre el nivel de experiencia de los individuos con la IA y las expectativas sobre su impacto. El esquema 2 representa este argumento.

Esquema 2



El esquema 2 sugiere que, en los individuos con baja experiencia, el miedo abstracto domina el escenario, lo que los impulsa a interpretar el temor como evidencia de peligro, lo que provoca un incremento en la probabilidad percibida. Entre los individuos con experiencia media, si bien el miedo concreto emerge, su efecto es ambiguo. Por último, entre los individuos con alta experiencia y conocimiento, el miedo concreto se encuentra presente, pero interpretado por medio de la comprensión técnica (marco epistémico más robusto), lo que provoca una disminución de la probabilidad percibida (al menos en el corto plazo). Esto permite afirmar que el efecto de la experiencia en los juicios de probabilidad es una relación no monotónica. Presentado de una manera alternativa (*vid.* tabla 3), el tipo de temor (*i.e.,* tipos de afecto) puede ser usado como predictor para cada categoría de respuesta. Retomo las categorías de respuesta establecidas en la tabla 1 y presento una propuesta interpretativa del miedo como predictor para las categorías de respuestas (*vid.* tabla 3).

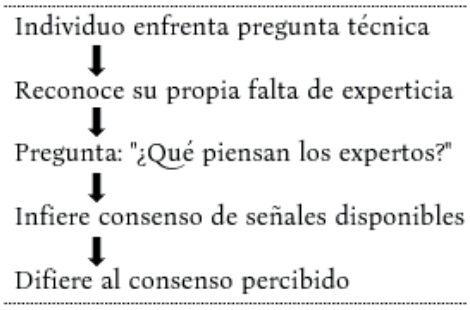
Tabla 3. El tipo de miedo predice categoría de respuesta

Categoría de Respuesta	Tipo primario de afecto	Evidencia
Y_1 "Ya es"	Impresión reciente + miedo vago	Colapsa ANI/AGI
Y_2 "Muy probable"	Alto temor concreto + extrapolación	Empleos, *deepfakes*
Y_3 "Algo probable"	Temor moderado	Cauteloso
Y_4/Y_5 Escépticos	Bajo temor o afecto contra-cultural	Dos subpoblaciones
Y_6 "No está seguro"	Bajo temor + humildad epistémica	Reconocimiento de ambigüedad

Nota. En base a la tabla 1.

Fundamentación de MP2. Cuando los individuos carecen de experticia para evaluar afirmaciones técnicas, dependen de información social (*cfr.* Bikhchandani *et al.*, 1992; Anderson y Holt, 1997), particularmente del consenso científico percibido (Bikhchandani *et al.*, 1992; Bromme y Goldman, 2014; Hendriks *et al.*, 2015). A continuación, el esquema 3 sintetiza este asunto.

Esquema 3



Nota. Literatura citada *supra*.

Ahora bien, la encuesta llevada a cabo por el Pew Research en 2025 compara la opinión del público con la de los expertos y revela una divergencia sistemática.

Tabla 4. Actitudes públicas vs. expertas hacia la IA

Dimensión	Público	Expertos	Brecha	Dirección
Perspectiva General				
Más emocionado que preocupado	11 %	47 %	+36 pts.	Expertos más positivos
Espera impacto positivo en EE. UU.	17 %	56 %	+39 pts.	Expertos más positivos
Espera beneficio personal	24 %	76 %	+52 pts.	Expertos mucho más positivos
Espera daño personal	43 %	15 %	-28 pts.	Público más pesimista
Preocupaciones Específicas				
Información inexacta	66 %	70 %	+4 pts.	Convergencia
Decisiones sesgadas	55 %	55 %	0 pts.	Convergencia perfecta
Pérdida de empleos	56 %	25 %	-31 pts.	Público mucho más preocupado
Pérdida de conexión humana	57 %	37 %	-20 pts.	Público más preocupado

Nota. Pew Research (McClain *et al.*, 2025).

La tabla 4 muestra que los expertos son sistemáticamente más optimistas (+36 a +52 puntos en perspectiva general); sin embargo, el público no percibe este optimismo. ¿Por qué? Una respuesta plausible podría ser que está operando un mecanismo de amplificación selectiva de señales, esto es, el público no accede al consenso experto directamente, sino a través de ciertos canales de filtrado. Se detectan, en el contexto de los datos, tres.

Figuras tecnológicas prominentes emiten una señal alarmista que luego se amplifica. Figuras de alta visibilidad que han emitido advertencias dramáticas,[10] convirtiéndose en altamente noticiables, citables/reproducibles

10. Figuras como Geoffrey Hinton y Yoshua Bengio alertaron (durante 2023) que "la IA podría llevar a la extinción" (Kleinman, 2023). En 2023, Elon Musk, Yoshua Bengio, Stuart Russell y una larga lista de figuras emblemáticas del nuevo sector tecnológico firmaron una carta abierta pidiendo pausar durante meses el entrenamiento de los grandes modelos (La Nación, 2023). La difusión fue tan amplia que Bengio escribió un artículo para el site del *Journal of Democracy* (Bengio, 2023).

con alto *feedback* y, además, contraintuitivas (llaman la atención porque son líderes tecnológicos advirtiendo sobre peligros en su propio campo), provocando que los expertos más optimistas reciban menos cobertura mediática, entre otras cosas, porque el optimismo es menos noticioso.

Koehler (2016) demostró experimentalmente que el "falso equilibrio" periodístico (dar igual peso a posiciones científicas mayoritarias y minoritarias) distorsiona la percepción pública del consenso. Aplicado al tema de este trabajo, se asume que a) la cobertura mediática frecuentemente presenta "beneficios de IA" vs. "riesgos de IA" como debate 50-50; sin embargo, b) el consenso experto (como lo demuestran los datos de la tabla 4) es más matizado (alta confianza en beneficios, menor confianza en riesgos existenciales); por tanto, c) resulta posible interpretar que aparece el sesgo de negatividad (*cfr.* Soroka y McAdams, 2015), es decir, que la evidencia psicofisiológica indica que las noticias negativas son más estimulantes y, también, *attention-grabbing* —para usar el término técnico por el cual son conceptualizadas— que las positivas, lo que ayuda a explicar el predominio de la negatividad en los *mass media*.

Los individuos, de acuerdo con el enfoque de la cognición cultural, acreditan o validan, selectivamente, información que se alinea con su cosmovisión cultural y desacreditan información que la amenaza (*cfr.* Kahan *et al.*, 2012). Llevado este razonamiento al tema que aquí interesa, se puede inferir que a) los individuos escépticos sobre los efectos de la IA (p. ej., preocupados por el poder corporativo, sensibles a los sesgos algorítmicos, a la pérdida de privacidad, etc.) atenderán, de manera preferencial, a las advertencias de críticos sobre la IA; mientras que b) los individuos optimistas (*v. gr.,* fe en la innovación, en las soluciones de mercado, etc.) mostrarán, preferentemente, mayor atención a los beneficios que promueven los defensores de la IA. Así, cabe suponer que ambos grupos encuentran incentivos para pensar algo como "Los expertos están de acuerdo conmigo" (sesgo de confirmación). La tabla 5 muestra una relación entre polarización política y expectativas sobre la IA.

Tabla 5. Identidad política y miedo a la IA

Grupo	2024 % "Asustado"	2025 % "Asustado"	Cambio (pp)	Dirección
Republicanos (R)	27 %	17 %	-10	Disminución
Demócratas (D)	20 %	24 %	+4	Aumento
Cambio	R +7 pts.	D +7 pts.	Inversión de 14 pts.	Reversión completa

Nota. YouGov (2024; 2025).

La tabla 5 documenta cómo, frente a la misma tecnología, en apenas un año, se observa una valencia emocional opuesta. Este patrón resulta consistente con la hipótesis de que, con el cambio de administración (la salida del presidente Biden y el regreso del presidente Trump), en enero de 2025 se produjo una modificación narrativa. Para los individuos identificados con el republicanismo, la IA comenzó, bajo esta interpretación, a operar en el contexto de una narración del tipo "La IA será liberada de la censura *woke*". Mientras que para los simpatizantes demócratas la narración sobre la IA entró en una zona de peligro del tipo "La IA será utilizada para la desinformación". Esta inversión de 14 puntos en un año sugiere que la IA se ha convertido en un marcador —como el cambio climático o las vacunas (*cfr.* Kahan *et al.*, 2012)— de identidad cultural/política. La afiliación grupal/identitaria parece funcionar como un predictor de actitudes hacia IA más que el conocimiento técnico. Si bien lo anterior puede interpretarse de manera alternativa,[11] la simetría inversa, la coincidencia temporal con la transición presidencial y su paralelismo con discusiones sobre polarización afectiva en el contexto político hacen que la interpretación aquí ofrecida resulte plausible.

Los anteriores mecanismos de filtrado pueden complementarse con una observación adicional. La tabla 6 muestra que, a pesar de la existencia de una brecha entre el público y los expertos de +52 puntos sobre el asunto del "beneficio personal", convergen en algunas preocupaciones específicas.

11. Entre aquellas interpretaciones alternativas se pueden mencionar dos: efectos de eventos específicos no partidistas entre 2024 y 2025 (p. ej., controversias sobre *deepfakes*, debates regulatorios, etc.) que afectaron diferencialmente a cada grupo por razones no relacionadas con identidad política o, también, cambios en la composición demográfica dentro de cada grupo (p. ej., entrada/salida de votantes jóvenes con actitudes distintas a la prevaleciente en grupos pre-existentes).

Tabla 6. Público vs. expertos: convergencia en peligros específicos			
Preocupación	**Público**	**Expertos**	**Cambio (pp)**
Información inexacta	66 %	70 %	+4
Sesgo algorítmico	55 %	55 %	0
Necesidad de regulación	71 %	57 %	-14

Nota. Pew Research (McClain *et al.*, 2025).

Valores positivos indican mayor prevalencia en expertos; valores negativos indican mayor prevalencia en el público. Por la naturaleza de la tabulación de los datos, estas diferencias asumen un carácter descriptivo.

Los expertos manifiestan estar más preocupados por algunos peligros específicos (p. ej., desinformación, sesgo) que por sobre otros (pérdida de empleos, extinción). El público recibe señales sobre qué asuntos preocupan a los expertos (transmisión precisa), pero no sobre cómo los expertos ponderan riesgos contra beneficios (transmisión imprecisa). El resultado es que el público piensa "los expertos están preocupados" (verdadero), entonces infieren que "los expertos piensan que la IA es negativa en términos netos" (algo que es falso).

Fundamentación de MP3. La teoría de los modelos mentales (*cfr.* Johnson-Laird, 1983) afirma que el razonamiento individual no está precedido por una rigurosa lógica formal, sino que se basa en representaciones internas de situaciones que se apoyan en la inferencia a través de la manipulación e inspección de información. Si bien los modelos mentales preservan una correspondencia entre la estructura relacional con la situación representada, resultan incompletos, es decir, solo capturan características sobresalientes, pero omiten detalles. En caso de situaciones ambiguas, los individuos pueden apelar a múltiples modelos alternativos, lo que supone una construcción inconsciente, es decir, automática o inadvertida por el individuo. En el caso que aquí interesa, los individuos, al enfrentarse con conceptos ambiguos (p. ej., "más inteligente que las personas"), pueden construir un modelo especificando: a) qué entidades se están comparando (¿sistemas ANI? ¿AGI futura? ¿ASI hipotética?), b) qué dimensión de la comparación resulta equivalente a "inteligencia", y c) qué constituye "más" (p. ej., ¿ligeramente mejor?, ¿vastamente superior?). Esto hace que diferentes individuos construyan diferentes modelos sin advertir que están respondiendo preguntas diferentes. Este asunto se profundiza cuando confrontamos al público con los expertos, ya que tienen diferentes estructuras conceptuales, es decir, no están procesando el mismo estímulo de manera diferente, están, en cambio, construyendo estímulos diferentes. La tabla 7 intenta capturar este crucial problema.

Tabla 7. Modelos mentales: novato vs. experto		
Constructo	**Modelo del novato**	**Modelo del experto**
"IA"	Entidad monolítica	Conjunto diverso de técnicas
"Inteligente"	Resultados impresionantes	Capacidades específicas medibles
"Más inteligente"	Mejor en cosas que valoro	Superior en *benchmarks* específicos
"Que las personas"	Que yo personalmente	Que un humano promedio/mejor/cualquiera.

La tabla 7 propone una interpretación en donde los expertos son capaces de distinguir la ANI (superior a los humanos en tareas específicas) de la AGI (cuya inteligencia resultaría similar a la de los humanos) de la ASI (una inteligencia superior a la de los humanos en todos los dominios). El público —es decir, los novatos— suele confundir estas distinciones y emplea el término "inteligente" para describir cualquier cosa impresionante. Con base en la propuesta de la tabla 7, analizo tres casos de modelos mentales.

Modelo 1. El individuo puede suponer que "más inteligente" equivale a "rendimiento superior en tareas observables". La forma en que llega a construir este modelo puede exponerse por medio del esquema 4.

Esquema 4

SI la IA puede hacer X mejor que humanos (donde X es impresionante)

ENTONCES la IA es "más inteligente que las personas"

Ejemplos de X:
- Jugar ajedrez perfectamente
- Resolver matemáticas complejas
- Escribir ensayos coherentes (ChatGPT)
- Generar imágenes realistas

Conclusión: La IA ya es más inteligente (porque ya hace estas cosas)

Los potenciales usuarios de este modelo tendrían un escaso conocimiento técnico, una alta exposición reciente al funcionamiento de la IA y funden las capacidades observables de la IA en tareas específicas con inteligencia general. Sería, bajo esta interpretación, el 13 % que respondieron "ya es más inteligente" (Y_1, tabla 1).

Modelo 2. El individuo podría pensar que "más inteligente" equivale a una AGI con capacidades generales similares a las de los humanos. El esquema 5 caracteriza la formación de este modelo mental.

Esquema 5

IA actual: Impresionante en dominios estrechos
Tendencia observada: Mejora rápida (GPT-2 → GPT-3 → GPT-4 → GPT-5)
Extrapolación: Si la tendencia continúa, entonces habrá una eventual inteligencia general

Horizonte temporal: No especificado, pero "en mi vida" (implícito)
Probabilidad: Moderada a alta (depende del *momentum* percibido)

Conclusión: Muy/algo probable que ocurra

Los usuarios de este modelo mental serían aquellos que poseen un conocimiento técnico intermedio, una continua exposición a las narrativas sobre el progreso de la IA y bastante confianza en los pronósticos de los expertos. Serían, conforme a la interpretación de los datos de la tabla 1, Y_2 e Y_3 (en conjunto el 47 % de los respondentes).

Modelo 3. Un individuo podría suponer que "más inteligente" consiste en una ASI.

Esquema 6

Distinguir:
- ANI (actual): Superioridad en tareas estrechas ✓ logrado
- AGI (futuro): Inteligencia general de nivel humano ✗ no logrado
- ASI (especulativo): Inteligencia radicalmente superior ✗ extremadamente difícil

"Más inteligente" = ASI (interpretación más alta)

Evaluación de barreras:
- Generalización: no resuelto
- Razonamiento de sentido común: brecha mayor
- Comprensión causal: los modelos actuales carecen
- Problema de alineamiento: sin solución de consenso

Conclusión: Improbable a corto/mediano plazo

El individuo típico de este modelo, es decir, que razonaría conforme al esquema 6, tendría un alto conocimiento técnico y estaría familiarizado con los desafíos de investigación en el campo de la IA (tanto si es optimista como pesimista). Conforme a la interpretación propuesta de la tabla 1, serían los respondentes ubicados en Y_4 e Y_5 (combinados conforman un 24 %).

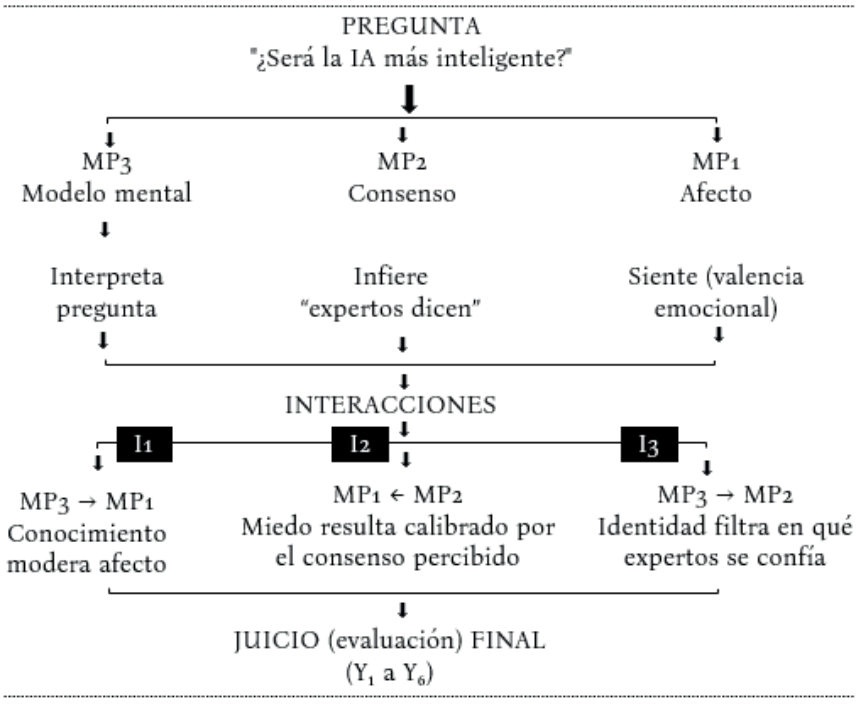
El 17 % de encuestados que seleccionó la opción "no está seguro" (Y_6) admite múltiples interpretaciones. Tres perfiles teóricos son compatibles con esta respuesta: (a) escaso conocimiento y baja exposición a discusiones sobre IA, (b) reconocimiento metacognitivo de la ambigüedad inherente a la pregunta o (c) incertidumbre genuina en individuos con conocimientos moderados que identifican tensión entre argumentos contradictorios.

Los patrones observados en otras secciones del análisis sugieren que la opción (b) es menos prevalente de lo que inicialmente podría esperarse. Los datos sobre expertos (*vid.* §.12) muestran que solo un 2 % considera que AGI existe actualmente y las diferencias público-expertos documentadas (*vid.* §.14-15) indican que el conocimiento técnico se asocia con juicios más calibrados y definidos, no con mayor suspensión del juicio. Esta evidencia es consistente con la hipótesis de que el grueso del 17 % refleja principalmente las opciones (a) y (c), es decir, personas con conocimiento insuficiente para formar una opinión o con conocimiento intermedio, pero que reconocen la presencia de argumentos contradictorios sin disponer de marcos conceptuales para resolverlos.

Bajo esta interpretación, la ambigüedad radical no operaría principalmente como reconocimiento reflexivo entre individuos informados, sino como incertidumbre estructural entre quienes carecen de modelos mentales técnicos suficientemente articulados (MP3, Modelo 3). Los datos sugieren que la familiaridad con la IA se asocia con reducción de esta incertidumbre, no porque resuelva la ambigüedad conceptual subyacente, sino porque facilita la adopción de algún modelo mental, aunque este pueda ser incompleto o técnicamente impreciso (como ilustra el 13 % que seleccionó la opción indicando que AGI ya existe).

Ahora expongo cómo los tres marcos de procesamiento propuestos (MP1, MP2 y MP3) interactúan. El esquema 7 representa tres tipologías diferentes de interacción (I1, I2 e I3).

Esquema 7



I1: El conocimiento modera la influencia afectiva, es decir, MP3 influye sobre MP1. Este mecanismo se puede describir del siguiente modo: el individuo con bajo conocimiento siente miedo frente a la IA e interpreta "Si estoy preocupado, debe ser peligroso". Por tanto, le adjudica una alta probabilidad percibida. Mientras que un

individuo con más conocimiento, frente al mismo temor, interpreta algo como "Mi temor es sobre problemas específicos, no amenaza existencial". Por tanto, la preocupación más calibrada provoca una probabilidad percibida distinta.

I2: La agrupación identitaria/cultural resulta en un filtro para seleccionar el consenso percibido. Esto quiere decir que, frente a las mismas declaraciones, observaciones y juicios expertos, existen interpretaciones opuestas (*vid.* tabla 5). El marcador identitario resulta capaz de filtrar selectivamente en MP2.

I3: La experiencia resulta un medio eficaz para recalibrar la relación entre la heurística afectiva (MP1) y los modelos mentales (MP3). De tal forma que la experiencia (mayor familiaridad) con la IA hace que los individuos perciban un menor miedo abstracto, pero un mayor temor concreto (MP1), mientras que los modelos mentales (MP3) permiten refinar, mediante el conocimiento técnico, las diferencias entre las AGI y las ANI.

RESULTADOS: UNA APROXIMACIÓN EXPLORATORIA

Ahora analizo los datos con base en los microfundamentos desarrollados en la sección anterior. Presento cinco patrones, cada uno ejemplificado mediante una pregunta teórica específica.

Heterogeneidad consistente. ¿La distribución de respuestas Y (*vid.* tabla 1) es consistente con ruido aleatorio o exhibe alguna estructura particular? Los datos muestran dos asuntos interesantes. En primer lugar, la distribución resulta marcadamente no uniforme; y, en segundo lugar, exhibe una distribución fuertemente desbalanceada. Así, sobresalen tres características relevantes.

Primera. El 13 % de los respondentes que seleccionó Y_1 ("ya es más inteligente") constituye una minoría sustantiva. Esta magnitud es consistente con la hipótesis de uso sistemático del modelo mental 1 (*vid.* §.8.), es decir, los respondentes realizaron una interpretación que refunde la AGI con la ANI. La comparación con las respuestas proporcionadas por los expertos refuerza esta interpretación, ya que solo el 2 % de los expertos cree que la AGI existe actualmente (una brecha de 11 puntos).

Segunda. La respuesta "no sé" (un 17 %) conforma un resultado muy elevado (*vid.* §.3), lo que sugiere que el reconocimiento de la ambigüedad puede deberse o etiquetarse en alguno de estos casos: a) los respondentes poseen escasa información (*i.e.,* algo así como "no sé suficiente"), b) reconocen la ambigüedad (es decir, se dicen a sí mismos "la pregunta no es clara") o c) se encuentran bajo genuina incertidumbre ("podría ir en cualquier dirección").

Tercera. Si se agrupan los respondentes Y_4 e Y_5 bajo la etiqueta "escépticos" (24 % combinado) y a los respondentes Y_1 e Y_2 en la categoría optimistas (38 % combinado), se observa que aproximadamente un cuarto de la muestra expresa escepticismo sustantivo sobre capacidades actuales o proyectadas de AGI. Este patrón es consistente con heterogeneidad genuina en los modelos mentales subyacentes, más que con una variación aleatoria en torno a una tendencia central.

En conjunto, estos datos sugieren que la heterogeneidad es sustancial y estructural. Esta configuración es compatible con la coexistencia de los tres modelos mentales propuestos (MP1, MP2, MP3), aunque los datos agregados no permiten asignar individuos específicos a modelos específicos ni estimar su prevalencia relativa con precisión.

La paradoja de la familiaridad. ¿La creciente familiaridad con IA reduce el miedo uniformemente o transforma el tipo de miedos? La predicción heurística-afectiva estándar (Sistema 1 del modelo dual) nos dice que si la familiaridad (con la IA) aumenta, entonces el miedo (hacia la IA) debería cambiar de manera uniforme. Los datos de la tabla 2 muestran, en cambio, que el miedo abstracto tiene una leve disminución, mientras que el temor concreto es grande. Este patrón es consistente con el MP1, es decir, a mayor familiaridad se evidencia menor miedo general. Además, el aumento de la familiaridad se asocia con un incremento en los temores específicos. Esto parece indicar que el miedo abstracto y específico tienen trayectorias opuestas, asunto que no resulta consistente con los modelos de afecto unitario clásico. Este análisis, a su vez, permite suponer que una mayor familiaridad con la IA no solo reduce el miedo, sino que lo transforma.

Brecha del consenso. ¿La percepción pública de la opinión experta coincide con la opinión experta real?

Tabla 8. Actitudes hacia la IA: público vs. expertos				
Dimensión	**Público**	**Experto**	**Cambio (pp)**	**Dirección**
Perspectiva general (% positivo)				
Más emocionado que preocupado	11 %	47 %	+36	Expertos más optimistas
Espera impacto positivo en EE. UU.	17 %	56 %	+39	Expertos más optimistas
Espera beneficio personal	24 %	76 %	+52	Expertos más optimistas
Espera daño personal	43 %	15 %	-28	Público más pesimista
Preocupaciones específicas (% preocupado)				
Información inexacta	66 %	70 %	+4	Convergencia alta
Sesgo algorítmico	55 %	55 %	0	Convergencia total
Pérdida de empleos	56 %	25 %	-31	Público más preocupado
Pérdida de conexión humana	57 %	37 %	-20	Público más preocupado

Nota. Pew Research (McClain *et al.,* 2025).

La tabla 8 documenta que la divergencia entre el público y los expertos es sistemática a nivel de la perspectiva general —destacando el mayor optimismo de los expertos respecto de los beneficios—, pero muestra convergencias en ciertos miedos o preocupaciones específicas. Esto es consistente con la hipótesis de que el público no accede al consenso experto de manera directa, sino a través de señales filtradas.

Bajo esta interpretación, los datos sugieren que el público sí infiere que los expertos están preocupados (consenso percibido), pero no infiere que, en conjunto, la mayoría de los expertos son netamente optimistas. Ello permite suponer que operan los canales de filtrado descritos en el MP2 del siguiente modo: el público adopta las preocupaciones específicas de los expertos (por ejemplo, sobre desinformación o sesgo), pero descarta —o no incorpora— el optimismo que estos manifiestan respecto de los beneficios que proyectan para la IA.

Esta lectura sugiere que el consenso percibido difiere del consenso real, fundamentalmente debido a los mecanismos de filtrado mediático y grupal/identitario. No obstante, no deben descartarse interpretaciones alternativas; entre ellas, conviene considerar diferencias en intereses y beneficios (económicos) entre expertos y público, la aplicación de distintas tasas de descuento temporal sobre beneficios futuros, o sesgos en la selección de expertos.

Aun así, la interpretación propuesta aquí parece robusta, en la medida en que resulta difícil explicar la convergencia selectiva sin atribuir algún rol a los procesos de filtrado.

Divergencia sobre la automatización laboral. ¿Los expertos y el público analizan, categorizan y describen los puestos de trabajo (ocupación) de manera diferente?

Tabla 9. Predicciones públicas vs. expertas: automatización laboral				
Ocupación	**Público %**	**Expertos %**	**Brecha**	**Interpretación**
Convergencia alta (brecha <10 pts)				
Cajeros	75 %	75 %	0	Automatización obvia
Periodistas	50 %	50 %	0	Observable (p. ej., GPT escribe)
Divergencia moderada (brecha >20 pts)				
Camioneros	33 %	62 %	+29	Expertos ven conducción automatizable
Maestros K-12	>50 %	<50 %	Inversión	Público ve "IA explica"; expertos ven necesidad de empatía
Médicos	>50 %	<50 %	Inversión	Público ve "IA diagnostica", expertos ven necesidad de juicio

Nota. Pew Research (McClain *et al.,* 2025).

La tabla 9 exhibe patrones diferenciados de convergencia y divergencia entre predicciones públicas y expertas sobre automatización laboral. Este patrón es consistente con la hipótesis que muestra que la convergencia se produce cuando la automatización resulta obvia o es observable, mientras que la divergencia se produce cuando lo que está en juego es la existencia de conocimiento técnico sobre el funcionamiento general de los algoritmos o fundamentos estructurales de la IA. La brecha de 29 puntos en el rubro "camioneros" ilustra bien este asunto. El público (el 33 %), se puede asumir, que parece razonar del siguiente modo: "Conducir es algo complejo; por tanto, se requiere inteligencia general". Dado que esta no está disponible, entonces lo más aconsejable es que no se automatice esta actividad o puesto de trabajo. Frente a este mismo asunto, los expertos (62 %) pueden, eventualmente, esgrimir un razonamiento muy diferente: "La conducción en autopistas es una tarea bien delimitada; con las técnicas actuales resulta posible y seguro automatizar el movimiento de los camiones en esas áreas" (lo que sí hay, pueden razonar los expertos, son barreras regulatorias o legales que retrasan o impiden la innovación/adopción).

Estos razonamientos hipotéticos son posibles porque operan con modelos mentales diferentes (*vid.* MP3). El público podría estar usando los modelos 1 y 2, que se pueden ejemplificar del siguiente modo: un individuo del público puede afirmar "conducir me parece difícil", añadiendo que se "requiere inteligencia general" y, dado que no hay una AGI, concluye que los camioneros (de carne y hueso, si se me permite esta expresión) resultan más seguros y confiables. Los expertos, en cambio, pueden razonar utilizando el modelo 3, pueden pensar "la conducción en autopistas es un problema de control estructurado" y agregar que "con la tecnología actual resulta posible, eficiente y, dentro de estándares mínimos, seguro"; concluyendo, así, que "la conducción de camiones es automatizable".

Sin dudas, este asunto puede interpretarse de otras maneras. Por ejemplo, por diferentes grados de tolerancia al riesgo, es decir, el público puede, antes de considerar una tecnología "automatizable", exigir estándares de seguridad más elevados que los expertos. También por un acceso diferencial a información vital; el público podría ignorar ciertos prototipos ya operativos en contextos delimitados (*i.e.,* asimetría informativa). O, simplemente, por intereses laborales divergentes o una aparente conducta prosocial basada en el egoísmo. Así, en este caso, el público podría expresar preferencias normativas ("no debería automatizarse") disfrazadas como predicciones técnicas ("no puede automatizarse"). Sin embargo, estas alternativas no explican fácilmente por qué ocurre convergencia en los ítems "cajeros" y "periodistas" mientras hay divergencia en "camioneros". Si fuera solo por tolerancia al riesgo o preferencias normativas, se esperaría una divergencia más uniforme. La especificidad del patrón (*i.e.,* convergencia donde la automatización es públicamente observable, divergencia donde requiere conocimiento técnico no evidente) apoya la interpretación basada en modelos mentales diferenciados.

La confianza calibrada a través de la experiencia. ¿La experiencia aumenta o disminuye la confianza en las capacidades de la IA? La predicción estándar es que la experiencia enriquece la familiaridad y esta produce una mayor confianza. Sin embargo, este quinto patrón muestra que la confianza no se derrumba ni se dispara, sino que se recalibra a través de la experiencia.

Tabla 10. Confianza en las capacidades de la IA

Dominio/Área/Campo específico	2024 %	2025 %	Cambio (pp)
Información precisa	43 %	48 %	+5
Decisiones sin sesgo	31 %	35 %	+4
Decisiones éticas	21 %	23 %	+2

Nota. YouGov (2024; 2025).

Los datos de YouGov sobre confianza en las capacidades de la IA (tabla 10) no muestran un colapso de la confianza, sino ajustes modestos pero sistemáticos. Entre 2024 y 2025, el porcentaje de estadounidenses que declara tener "mucha" o "bastante" confianza en que la IA provea información precisa pasa de un 43 % a un 48 %, en que tome decisiones sin sesgo de un 31 % a un 35 % y en que tome decisiones éticas de un 21 % a un 23 %; mientras que las respuestas "no sabe" disminuyen en los tres casos. En otras palabras, este patrón sugiere que la creciente familiaridad con la IA se asocia con una reducción de la suspensión del juicio y desplaza a una fracción de los indecisos hacia posiciones definidas de confianza o desconfianza, sin que ello implique una transformación radical en las actitudes agregadas.

Por otra parte, estos datos revelan que, si bien la desconfianza sigue siendo elevada, sobre todo en materia de decisiones éticas y sesgos, coexiste con un ligero aumento de la confianza declarada. Esta combinación, en términos generales, resulta más compatible con un proceso de calibración gradual que con una dinámica lineal de aclimatación optimista (del tipo "a más experiencia, más confianza") o de alarma creciente (de la modalidad "a más experiencia, más miedo"). Los datos sugieren que la experiencia con la IA permite identificar ámbitos en los que los encuestados están dispuestos a concederle cierta competencia, pero, al mismo tiempo, existe la percepción de riesgos específicos.

Esta dinámica, vista desde el MP3, sugiere que la experiencia no opera solo como acumulación de información, sino como reorganización de modelos mentales. Los individuos aprenden a distinguir tareas en las que la IA puede "funcionar" razonablemente bien de aquellas en las que su intervención se percibe como especialmente peligrosa o normativamente problemática. La familiaridad reduce el "no sé", pero no disuelve el miedo; parece, en todo caso, redistribuirlo hacia ámbitos específicos (trabajo, privacidad, manipulación, *deepfakes*). Esta mezcla de confianza moderada y miedos focalizados constituye, desde la perspectiva del análisis, una zona gris, ya que los datos no validan una aceptación cándida, pero tampoco la existencia de un miedo social creciente. No se puede descartar que estén operando otros factores, tales como una cobertura mediática menos alarmista en 2025. También que las actitudes iniciales de los respondentes en las diferentes cohortes sean diferentes o, incluso, cambios en referentes cognitivos acerca de la IA. Sin embargo, la consistencia direccional de los cambios (aumento modesto pero uniforme en las tres dimensiones) y su coherencia con patrones de familiaridad-especificidad documentados en las otras tablas hacen que la interpretación de calibración mediante reorganización de modelos mentales resulte atendible.

CONCLUSIÓN

En este artículo argumenté que las expectativas públicas sobre capacidades de la IA emergen de la interacción de heurísticas afectivas moduladas por familiaridad (MP1), del anclaje en el consenso experto percibido (MP2) y de la construcción de modelos mentales (MP3).

El análisis exploratorio de datos de YouGov (2024, 2025) y Pew Research (2023-2025) reveló cinco patrones convergentes: a) la existencia de una heterogeneidad sustancial, lo que sugiere la presencia de múltiples modelos mentales; b) la presencia de la paradoja de la familiaridad, hallazgo que apoya la idea que los afectos tienen modalidades diferentes, no solo intensidades variables; c) la existencia de una brecha de consenso, evidencia que apoya la idea de consenso filtrado; d) la presencia de una divergencia en cuanto a la idea de la automatización (laboral), lo que sugiere la existencia operativa de diferencias entre modelos mentales; y, finalmente, e) una recalibración de la confianza a través de la experiencia, lo que indica que esta última no produce ni un optimismo acrítico ni un rechazo uniforme, sino una combinación de confianza moderada y miedos focalizados.

Este análisis se construyó a partir de datos secundarios, lo que supone no disponer de microdatos brutos; por tanto, no fue posible estimar asociaciones individuales, interacciones entre mecanismos o controles sistemáticos. El estudio sobre Estados Unidos puede ser una doble limitación; en primer lugar, estos datos pueden estar exagerando patrones exclusivamente domésticos; y, por otro, la propia heterogeneidad subnacional de EE. UU. podría modular los resultados. Por último, la comparabilidad de los datos pudo verse afectada por cambios semánticos (como el registrado en la Encuesta Chapman) o por modificaciones en los marcos de referencia (las diferentes encuestas de Pew Research).

Así, cabe señalar que los patrones presentados tienen un carácter exploratorio, en tanto su organización no demuestra ni establece relaciones causales. Finalmente, este trabajo ofrece un mapa exploratorio de cómo las personas imaginan, sienten y discuten la IA; por ello, los hallazgos aquí presentados deben leerse como indicios sugestivos, no como evidencias firmes. En esta línea, una prolongación natural del análisis consistiría en examinar la relación entre miedo y preferencias regulatorias incorporando medidas de confianza institucional, con el fin de evaluar si (y eventualmente cómo) la desconfianza hacia el Estado modera la demanda de regulación.

REFERENCIAS BIBLIOGRÁFICAS

Anderson, L. R. y Holt, C. A. (1997). Information cascades in the laboratory. *American Economic Review, 87*(5), 847-862. https://doi.org/10.1016/S0167-4870(00)00025-8

Bengio, Y. (2023). AI and Catastrophic Risk. *Journal of Democracy*. https://bit.ly/4iJMPAn

Bikhchandani, S., Hirshleifer, D. y Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of Political Economy, 100*(5), 992-1026. https://doi.org/10.1086/261849

Bromme, R. y Goldman, S. R. (2014). The public's bounded understanding of science. *Educational Psychologist, 49*(2), 59-69. https://doi.org/10.1080/00461520.2014.921572

Chi, M. T., Feltovich, P. J. y Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*(2), 121-152. https://doi.org/10.1207/s15516709cog0502_2

Evans, J. St. B. T. y Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8*(3), 223-241. https://doi.org/10.1177/1745691612460685

Faverio, M. y Tyson, A. (2023). *What the data says about Americans' views of artificial intelligence.* Pew Research Center. https://bit.ly/47oUvnG

Finucane, M. L., Alhakami, A., Slovic, P. y Johnson, S. M. (2000). The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making, 13*(1), 1-17. https://doi.org/10.1002/(SICI)1099-0771(200001/03)13:1

Fischhoff, B., Slovic, P., Lichtenstein, S., Read, S. y Combs, B. (1978). How safe is safe enough? A psychometric study of attitudes towards technological risks and benefits. *Policy Sciences, 9*(2), 127-152. https://doi.org/10.1007/BF00143739

Gigerenzer, G. y Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103*(4), 650-669. https://doi.org/10.1037/0033-295X.103.4.650

Hendriks, F., Kienhues, D. y Bromme, R. (2015). Measuring laypeople's trust in experts in a digital age: The Muenster Epistemic Trustworthiness Inventory (METI). *PLOS ONE, 10*(10). https://doi.org/10.1371/journal.pone.0139309

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.

Kahan, D. M., Jenkins-Smith, H. y Braman, D. (2011). Cultural cognition of scientific consensus. *Journal of Risk Research, 14*(2), 147-174. https://doi.org/10.1080/13669877.2010.511246

Kahan, D. M., Peters, E., Wittlin, M., Slovic, P., Ouellette, L. L., Braman, D. y Mandel, G. (2012). The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nature Climate Change, 2*(10), 732-735. https://doi.org/10.1038/nclimate1547

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Kleinman, Z. (2023). "Yoshua Bengio, el «padrino» de la IA, se siente «perdido» en lo que respecta a su trabajo." BBC. 31 de mayo. https://bit.ly/3LoZD2j

Koehler, D. J. (2016). Can journalistic "false balance" distort public perception of consensus in expert opinion? *Journal of Experimental Psychology: Applied, 22*(1), 24-38. https://doi.org/10.1037/xap0000073

Krosnick, J. A., Holbrook, A. L., Berent, M. K., Carson, R. T., Hanemann, W. M., Kopp, R. J., Mitchell, R. C., Presser, S., Ruud, P. A., Smith, V. K., Moody, W. R., Green, M. C. y Conaway, M. (2002). The impact of "no opinion" response options on data quality: Non-attitude reduction or an invitation to satisfice? *Public Opinion Quarterly, 66*(3), 371-403. https://doi.org/10.1086/341394

La Nación. (2023). "Elon Musk y miles de expertos pidieron hacer una "pausa" en los avances de la Inteligencia Artificial durante 6 meses". 30 de marzo. https://bit.ly/4oy1Irt

Lewandowsky, S., Gignac, G. E. y Vaughan, S. (2013). The pivotal role of perceived scientific consensus in acceptance of science. *Nature Climate Change, 3*(4), 399-404. https://doi.org/10.1038/nclimate1720

McClain, C., Kennedy, B., Gottfried, J., Anderson, M. y Pasquini, G. (2025). *How the U.S. Public and AI Experts View Artificial Intelligence*. Pew Research Center. https://bit.ly/47HIKYj

Oreskes, N. (2004). The scientific consensus on climate change. *Science, 306*(5702), 1686. https://doi.org/10.1126/science.1103618

Schwarz, N. y Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology, 45*(3), 513-523. https://doi.org/10.1037/0022-3514.45.3.513

Siegrist, M., Earle, T. C. y Gutscher, H. (2005). *Trust in risk management: Uncertainty and scepticism in the public mind*. Earthscan.

Sjöberg, L. (2000). Factors in risk perception. *Risk Analysis, 20*(1), 1-12. https://doi.org/10.1111/0272-4332.00001

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*(1), 3-22. https://doi.org/10.1037/0033-2909.119.1.3

Slovic, P., Finucane, M. L., Peters, E. y MacGregor, D. G. (2004). Risk as analysis and risk as feelings: Some thoughts about affect, reason, risk, and rationality. *Risk Analysis, 24*(2), 311-322. https://doi.org/10.1111/j.0272-4332.2004.00433.x

Soroka, S. y McAdams, S. (2015). News, politics, and negativity. *Political Communication, 32*(1), 1-22. https://doi.org/10.1080/10584609.2014.881942

Tversky, A. y Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131. https://doi.org/10.1126/science.185.4157.1124

YouGov. (2024). "Artificial Intelligence". https://bit.ly/47V9IM7

YouGov. (2025). "Artificial Intelligence". https://bit.ly/4nnw4eQ

Yudkowsky, E. y Soares, N. (2025). *If anyone builds it, everyone dies. Why superhuman AI would kill us all*. Little, Brown & Company.